WitFoo

STRATEGIC WHITE PAPER

The AI cost collision in cybersecurity

Why empathetic processing becomes the economic imperative of the next decade

Charles D. Herring

October 2026

WITFOO, INC.

© 2026 WitFoo, Inc. CC BY 4.0

Cite as: Herring, C. D. (2026). The AI cost collision in cybersecurity. WitFoo, Inc.

CONTENTS

Contents

Foreword

Executive summary

Part I · The AI cost collision

1. How we got here — the AI optimism era
2. The supply wall — chips, buildings, electrons
3. Token economics — what frontier inference actually costs
4. Why security sits in the crosshairs
5. What the agencies now advise

Part II · Why AI-first SOCs break at scale

6. The feed-everything-to-AI pattern
7. The three taxes — redundancy, noise, retrieval
8. A worked cost example

Part III · Empathetic processing

9. What empathetic processing means
10. ProtoGraph deduplication
11. The pre-AI curation discipline
12. Temporal Link Analysis — from signals to structured incidents
13. Where AI earns its cost

Part IV · The economic math

14. Data volume reduction
15. Token savings
16. Storage, compute, and transit savings
17. The WitFoo pricing model

Part V · The strategic outlook

18. The AI supply and demand curve
19. Compounding advantage
20. Recommendations for 2026 through 2029

Conclusion

How this paper was produced

About WitFoo

Appendix A · WitFoo pricing reference

Appendix B · References

Dollar figures are US dollars unless marked NZD.

FOREWORD

Every technology wave ends in arithmetic.

In cybersecurity, we are entering the arithmetic stage of the AI wave. The industry has spent three years telling itself a story: that large language models will triage every alert, investigate every anomaly, summarize every incident, and eventually replace entire tiers of the SOC. Vendor roadmaps have been built on this promise. CISOs have been told to AI-enable their operations. CFOs have signed budgets assuming the economics will work out.

They may not.

The AI compute market is on a collision course with physical reality. Chip fabrication capacity is rationed. Data center construction has multi-year lead times. Power grids built for the last century cannot serve this century's inference demands. Token prices at the frontier have stopped falling in step with model quality gains. The Uptime Institute now identifies power (not talent, not innovation, not regulation) as the single defining constraint on data center growth.

In security, this collision will arrive sooner than anywhere else. Few other domains generate data volumes this large or demand per-event analysis this deep. The "feed everything to the LLM" architecture that product marketing loves is an economic dead end, and the smarter practitioners already know it.

Since the first draft of this paper, the agencies that defend national networks have done the same arithmetic. In October 2026 the national cyber security agencies of Australia, Canada, New Zealand, the United Kingdom and the United States published joint guidance on managing AI token consumption. They put cyber defense, incident response and threat intelligence at the front of the queue for tokens, they warn about attacks designed to run up an AI bill, and they suggest preprocessing and filtering data before it reaches a model. None of it names a product, and none of it should. But it moves this conversation from the vendor's pitch to the operator's checklist, which is where it belongs.

This paper makes a different case. It argues that the future of AI in cybersecurity is not more AI, but better-prepared data. It introduces empathetic processing, a security-aware, deterministic pre-AI layer that strips noise, deduplicates evidence, and normalizes context, so that the expensive parts of the AI pipeline touch only the signals that actually warrant it. It presents the economic math, and it argues that WitFoo's per-appliance, unlimited-data pricing model is uniquely positioned to compound in value as AI costs continue to rise.

We are not anti-AI. We are pro-efficient-AI. The difference, as this paper will show, is measured in tens of millions of dollars per enterprise over the next five years, and increasingly in whether your AI-enabled SOC can get inference capacity at all.

Charles Herring, Chairman and co-Founder, WitFoo

Christchurch, New Zealand · October 2026

EXECUTIVE SUMMARY

The arithmetic of AI in security.

The cybersecurity industry is building its next decade of operations on an economic assumption that is beginning to crack: that AI compute will keep getting cheaper faster than security data volumes grow. It will not.

Three forces are converging

AI compute demand is growing exponentially across training, inference, and — most aggressively — the newly dominant category of agentic AI, which issues many LLM calls per task. Supply-side constraints are biting: GPU fabrication capacity is rationed, data center construction has two- to three-year lead times, and the US power grid cannot absorb the buildout hyperscalers are planning. Gartner forecasts that power shortages will restrict roughly forty percent of AI data centers by 2027. And security data volumes continue to grow thirty to forty percent year over year, driven by cloud workloads, endpoint telemetry, and identity signals with no upper bound in sight.

The arithmetic is brutal

A security operation ingesting five hundred gigabytes of data per day and routing a meaningful fraction through frontier LLMs faces five- to six-figure monthly token bills at current pricing — and frontier token prices have stopped falling. Organizations architecting around the assumption that AI will simply handle it are building balance-sheet time bombs. They are also exposing themselves to a second risk that is harder to quantify: inference capacity rationing, as hyperscalers begin prioritizing their highest-paying enterprise customers for constrained GPU minutes.

The agencies have named the risk

In October 2026 the national cyber security agencies of Australia, Canada, New Zealand, the United Kingdom, and the United States published joint guidance, *Managing AI token consumption for operational resilience*. It treats token consumption as a resilience risk as well as a cost: unpredictable bills, token exhaustion when quotas run out, and denial-of-wallet attacks on pay-per-use pricing. It puts cyber defense, incident response, and threat intelligence in the top tier for token availability, and it suggests preprocessing and filtering data before it reaches a model. Empathetic processing is that discipline, built for security data.

The answer is not less AI. It is better-curated data.

WitFoo calls this empathetic processing: a security-aware, deterministic pre-AI layer that understands the relationships between clients, servers, users, files, and telemetry. It removes more than ninety percent of data volume through deduplication and normalization before any token leaves the customer’s environment. That is not sampling: nothing security-relevant is discarded, and every record that reaches downstream systems points back to its raw evidence. AI is then applied surgically — to the two to ten percent of signals that actually warrant expensive reasoning — rather than as a firehose against an ocean.

WitFoo Conductor is this layer. Deployed close to data sources, Conductor uses ProtoGraph deduplication, a security-aware reduction technique, together with Adaptive Parsing, a zero-maintenance signal normalization approach. The output feeding any downstream system — SIEM, AI SOC platform, data lake, or WitFoo Analytics — is cleaner, smaller, and semantically richer than the input.

The WitFoo pricing model compounds this advantage

Unlike SIEMs, data pipelines, and AI SOC platforms that charge by data volume, daily ingest, per alert, or per token, WitFoo appliances are priced per unit with unlimited data rates within each appliance's capacity envelope. A Conductor Pro costs forty thousand US dollars per year regardless of how much data it processes inside that envelope. Very high volumes may call for additional Conductors, but the per-appliance price never changes, and appliance count grows far more slowly than volume does. As customer data expands, WitFoo costs move in step-functions of appliance count rather than linearly with GB/day. As AI per-token prices hold or rise under supply pressure, WitFoo customers are shielded from that too.

Combined economic impact

A typical mid-market enterprise running a Splunk-centric SIEM with AI triage layered on top could spend one and a half to three million US dollars annually on the combination of SIEM licensing, storage, and AI inference. A WitFoo-fronted architecture using Conductor Pro to reduce ingest and route only high-value signals to AI delivers more than ninety percent reduction in SIEM ingest volume, roughly an eighty-five percent reduction in AI inference and retrieval spend in this paper's worked example, fixed and predictable infrastructure costs that do not scale with data growth, and a natural land-and-expand path toward replacing the SIEM entirely over time.

Strategic outlook

WitFoo's view is that the AI supply and demand curve will keep tightening through at least 2028 to 2029. Enterprises that lock in AI-heavy architectures today — with per-token or per-alert cost exposure — will face progressively painful renewals as capacity tightens and pricing stabilizes above where the industry expected it to settle. Enterprises that architect around empathetic processing will experience the opposite: compounding cost advantage as industry pricing pressure rises.

The next four to six years will separate security programs that treat AI as a precision tool from those that treat it as a firehose. Only the former will be affordable — and as capacity tightens, only the former will even be buildable.

PART I

The AI cost collision

Three macroeconomic forces are moving toward each other on timelines that do not align. This section lays them out: the trajectory of AI compute demand, the supply-side constraints that will not flex in time, and the specific way security workloads sit in the worst possible part of the curve. It closes with what five national cyber security agencies now advise.

1 · How we got here — the AI optimism era

For the last three years, the default assumption inside security operations has been that AI is a rising tide. Model quality has improved monthly. Per-token prices for mid-tier models have dropped dramatically — GPT-4o-mini, Claude Haiku, Gemini Flash, DeepSeek, and several open-weight alternatives now run at a small fraction of 2023 frontier pricing. Every SIEM, XDR, and emerging AI SOC vendor has launched an AI feature: copilots for triage, natural-language search, automated summarization, agentic investigation.

This has produced a predictable architectural pattern. Feed as much data as possible into the LLM. Let it reason across alerts, endpoints, cloud workloads, identities. Let it generate incident narratives, draft playbooks, write customer communications. The more context the model has, the better the output — and context is cheap, so pour it in.

That pattern works beautifully in a demo. In a Tier-1 enterprise running twenty terabytes of raw security telemetry per day, it breaks.

The optimism had three assumptions

First, that compute capacity was infinite. Cloud providers have spent a decade training the industry to believe that any scale was available at some price. If your workload grew tenfold, you paid tenfold, but you could always buy the capacity.

Second, that per-token prices would trend toward zero. Every six months, a new model arrived that was cheaper and better than the previous generation. Extrapolation said frontier inference would be commoditized within two to three years.

Third, that agentic workflows would amortize cheaply. A well-designed agent makes five, ten, fifty LLM calls per investigation — but if each call costs fractions of a cent, the total is still affordable.

All three assumptions are now under pressure. Compute is suddenly scarce; frontier prices are showing a floor; and agentic workflows, at the volumes security operations actually produce, sum to numbers that no CFO wants to see on a renewal.

> **The pivot in 2026**
>
> The same research that two years ago projected a token-price race to the bottom now projects a price floor at the frontier, rationed capacity at hyperscalers, and on-site power generation deals because the grid cannot keep up. What changed is not model capability. What changed is that reality caught up with the buildout plan.

2 · The supply wall — chips, buildings, electrons

Three supply-side constraints govern how fast AI inference capacity can grow. Each has a different time constant, and the slowest one sets the pace for the entire system.

Silicon

Advanced accelerator chips — NVIDIA Blackwell and Hopper, Google TPU, AMD MI300 and successors — are fabricated at a small number of leading-edge foundries, overwhelmingly by TSMC in Taiwan. Capacity is allocated in advance, often years ahead of delivery. Hyperscalers book the bulk of capacity for their own internal inference needs and for large enterprise customers. Benchmark data from the AI compute markets shows Blackwell-class GPU rental indices surging twenty-plus percent in a single month in early 2026 while the older Hopper-class inventory barely moved — the classic signature of a constrained market at the top of the capability curve.

Buildings

Even when silicon is available, it has to be placed somewhere. Data centers take two to three years to build from greenlight to commission. CBRE reports North American data center vacancy rates fell to a record-low 1.6 percent in 2025, with large-scale (10 MW plus) rental rates up nineteen percent year over year. Over seventy-four percent of new supply is pre-leased before commissioning. The traditional hubs — Northern Virginia, Silicon Valley, London — are power-constrained, and the industry has begun routing capital to wherever electrons are available. Proposed new data centers increasingly face local political opposition, which can block, slow, or limit build-outs — particularly for the largest facilities.

Electrons

Power is the true bottleneck. The International Energy Agency projects global data center electricity consumption reaching roughly 1,100 terawatt-hours in 2026 — comparable to the entire electricity consumption of Japan. The Uptime Institute now identifies power as the single defining constraint on data center growth globally, and predicts that AI-specific data center power load will hit ten gigawatts by the end of 2026 — a figure that reflects what the grid can deliver, not what the industry wants to deploy. Gartner forecasts that power shortages will restrict roughly forty percent of AI data centers by 2027.

New grid capacity is slow. A transmission line can take seven to ten years from planning to energization. Large baseload generation — nuclear small modular reactors, gas turbines, large-scale renewables plus storage — takes comparable timeframes. In the interim, hyperscalers are doing something that would have been

unthinkable three years ago: striking direct deals with energy majors for dedicated on-site generation. Microsoft's 2026 negotiation with Chevron for a natural gas facility to directly power a Texas data center is not an outlier; it is the new pattern. When the grid cannot deliver, data centers are building around it.

40%

of AI data centers are forecast to be restricted by power shortages by 2027 — Gartner

The compounding effect

These three constraints do not just add; they multiply. A project stalls on permitting, which delays data center commissioning, which pushes silicon deliveries out, which rations inference capacity for the workloads that actually made it online. The system has no slack. And security data — which grows faster than almost any other category of enterprise data — sits downstream of all three bottlenecks.

> **The planning gap**
>
> The technology sector builds on product cycles measured in months. The energy sector builds on infrastructure cycles measured in decades. The IEA's analysis is explicit about this mismatch, and it is not a temporary disruption — it is a structural feature.
>
> For security leaders, the implication is simple: capacity may not be contractable at the prices you modeled, and in some regions may not be contractable at all, through the end of the decade.

3 · Token economics — what frontier inference actually costs

To reason about AI cost in security, you have to reason about tokens. Every piece of data sent to an LLM — every alert, every log snippet, every context window of related events — consumes input tokens. Every word the model produces in response consumes output tokens. The two are not priced equally.

The input-output asymmetry

Across every major frontier lab, output tokens cost three to eight times more than input tokens. The reason is computational: generating tokens autoregressively, one step at a time with full context attention at each step, is fundamentally more expensive than embedding inputs in parallel. In the current pricing landscape, output tokens account for most of the bill on any generation-heavy workload — including security summarization, narrative generation, and agentic investigation.

The 2026 frontier pricing landscape

The table below captures representative published pricing as of early 2026. Exact numbers shift month to month, but the structural pattern — a 3-8× output premium, and a roughly 100× spread from cheapest-throughput to frontier-reasoning models — has held steady.

Model tier (representative)	Input / million tokens	Output / million tokens	Output / input ratio
Ultra-light (Nano / Flash-Lite / Mistral Small)	$0.10 – $0.15	$0.30 – $0.60	3 – 4×
Mid-tier (GPT-4.1, Sonnet-class)	$2.00 – $3.00	$8.00 – $15.00	4 – 5×
Frontier reasoning (Opus, GPT-5, Gemini Pro)	$5.00 – $15.00	$25.00 – $75.00	5×
Premium reasoning (GPT-5.2 Pro class)	$21.00	$168.00	8×

Source: aggregated from Silicon Data, AI Cost Board, Price Per Token, and CostGoat January–April 2026 snapshots. Pricing is representative and changes frequently.

The price-floor phenomenon

Between 2023 and 2025, frontier model pricing fell roughly thirty to fifty percent per year. That trajectory has slowed materially. New frontier models launch at similar pricing to the previous generation; the savings arrive only as those models move from flagship to mid-tier. For any security workload that needs best-available reasoning — high-stakes investigation, complex correlation, polished analyst output — the price is stabilizing, not dropping.

There are three reasons. First, the industry has hit the limits of easy efficiency gains; further reductions require new hardware generations. Second, capacity constraints give providers less incentive to cut prices — why compete on price when demand already exceeds supply, especially when AI providers themselves are under increasing pressure to show a path to profitability? Third, the compute cost of frontier reasoning models is structurally higher than first-generation GPT-class models because of larger parameter counts, longer thinking traces, and multi-pass verification.

The cache and batch discounts

Providers have introduced prompt caching (50 to 90 percent discount on cached repeated context) and batch APIs (typically 50 percent discount for non-real-time workloads). These are genuinely useful optimizations — and security workloads with repeated system prompts benefit. But they do not change the fundamental supply/demand dynamic. They simply allow providers to price-discriminate between latency-sensitive and latency-tolerant workloads. Most SOC use cases are latency-sensitive by nature: you cannot batch an active incident.

> **What the token numbers really tell us**
>
> The 3–8× output versus input token premium is the most important structural fact in AI economics right now. Workloads that generate a lot of output — narrative summaries, investigation reports, reasoning traces — are several times more expensive per “round trip” than workloads that classify or extract. Agentic workflows that iterate through many output-heavy turns compound this premium quickly.

4 · Why security sits in the crosshairs

Every enterprise function that wants to adopt AI faces the supply and pricing realities described above. But security faces them harder than most. Four structural factors put security workloads at the wrong end of every cost curve.

Data volume

Enterprise security data — firewall logs, endpoint telemetry, cloud audit trails, identity events, DNS, network flows, vulnerability scans, email metadata — routinely runs from hundreds of gigabytes to multiple terabytes per day. Splunk customers commonly report ingesting 500 GB to 5 TB daily. This is orders of magnitude larger than most other AI workload inputs. If even a modest fraction of this data flows through an LLM, the token bill becomes the dominant operating cost line item.

Per-event analysis demand

Unlike workloads where a single document or query needs AI reasoning, security analysis wants per-event reasoning. Every alert needs triage. Every anomaly needs context. Every potential incident needs correlation across sources. The natural instinct is to give the LLM the full event plus its surrounding context — which multiplies input tokens by an order of magnitude per event.

Agentic amplification

Autonomous SOC platforms — the category that D3 Security, Dropzone, Intezer, and others now compete in — produce dramatic results by running multi-step agentic investigations. An agent investigating a single alert may make five to fifty LLM calls: gather context, enrich with threat intel, correlate with other alerts, hypothesize attack paths, generate an incident narrative. Each call is modest. The aggregate, across an enterprise producing tens of thousands of alerts per day, is not.

Dropzone's early-2026 published pricing of roughly $36,000 per year for 4,000 investigations implies about $9 per investigation — which is the all-in cost of the AI, the platform, and the engineering behind it. D3 Morpheus quoted $0.27 per alert at enterprise scale at the same time. These are reasonable numbers for curated, high-signal alert streams. They are not reasonable numbers if you are feeding an unfiltered firehose of hundreds of thousands of daily alerts.

Retention and compliance pressure

Regulated industries — finance, healthcare, government, critical infrastructure — have retention and replay requirements that conventional cost optimization cannot satisfy. You cannot sample security data the way observability tools sample traces, because a sampled-out event is the event an investigator would have needed six months later. AI can help with retention-heavy analysis (historical correlation, threat-hunt lookbacks), but only if the data being retained is small enough to afford.

> *Cybersecurity expert.*
>
> *"Log collection and storage is often one of the largest ticket items in security operations budgets. These costs are ballooning now that everyone wants to read all their logs with AI. It makes sense to preprocess log data to convert it into higher level information that AI models can consume more rapidly and efficiently, and doing so frees up token budget that is needed for other critical security objectives."*
>
> **Tom Cross**, cybersecurity expert and WitFoo advisor

> **The compounding effect**
>
> Security is simultaneously the most data-intensive AI workload, the most per-event-hungry, the most agentic-friendly, and the most retention-bound of any enterprise domain. All four factors push cost in the same direction, higher. This is why the SIEM-plus-AI architecture that the industry has embraced for the last three years is a fiscal dead end unless something reduces the data volume before it touches the model.

That "something" is what the next three parts of this paper are about.

5 · What the agencies now advise

In October 2026 the question this paper asks stopped being a vendor's question. The Canadian Centre for Cyber Security and its partner agencies (the Cybersecurity and Infrastructure Security Agency in the United States, the Australian Signals Directorate's Australian Cyber Security Centre, and the National Cyber Security Centres of New Zealand and the United Kingdom) published joint guidance, *Managing AI token consumption for operational resilience* (ITSM.00.125). Its subject is how organizations understand, manage, and reduce the tokens their AI services consume, and why that matters for resilience as much as for cost. The guidance names no vendors and no products. It describes a discipline.

The same drivers this paper describes

The guidance lists the conditions under which token usage and cost increase: larger inputs and outputs; large reference datasets and connector payloads that are not cached or optimized; additional data and conversation history included with requests; multi-step workflows; repeated requests that process the same data multiple times; integrated systems that retrieve or process data across connected platforms; and agentic workflows that compound costs, "with each step reprocessing prior content and tool outputs." Read against Part II of this paper, the list is the redundancy tax, the noise tax, and the retrieval tax, described by the agencies whose job is to keep critical services running.

Risks beyond the bill

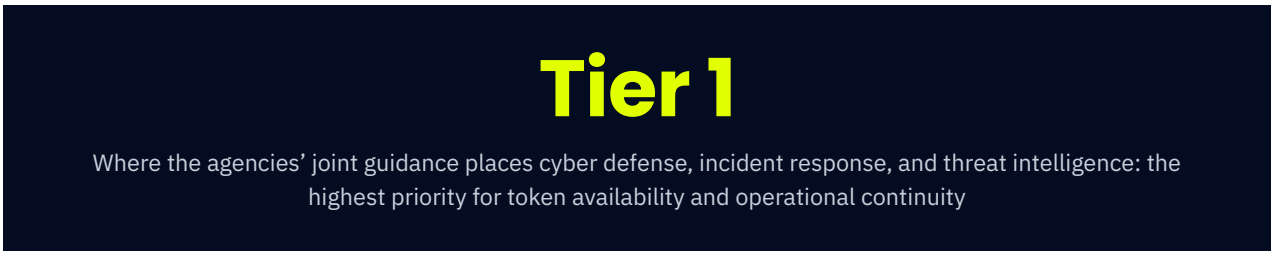
The guidance treats rising and unpredictable cost as only the first risk. The others are operational: lower-value activity consuming capacity that critical operations need; limited visibility into how tokens are used; slower and less consistent results as token counts grow; and token exhaustion, when an organization reaches its quotas, rate limits, or budget thresholds and its AI capabilities degrade or stop. It also names two forms of

deliberate abuse. In a denial-of-tokens attack, threat actors generate excessive usage to exhaust allocations or trigger rate limits. In a denial-of-wallet attack, they exploit pay-per-use pricing to drive up costs, for example with complex inputs that demand long outputs, or with loops that feed a tool's output back into it.

Cyber defense sits in the top tier

To manage those risks, the guidance recommends a tiered governance model that separates token allocations by business impact. Tier 1, mission critical, includes cyber defense, incident response, and threat intelligence: services that "should receive the highest priority for token availability and operational continuity." Tier 2 covers essential business processes. Tier 3 covers everyday productivity work that may face usage limits when demand is high.

The implication for security leaders is direct. The workload the agencies most want protected is the one this paper identifies as the most data-intensive, most per-event, most agentic, and most retention-bound in the enterprise. A SOC that spends its token budget proving that duplicate and low-value events were benign is spending Tier 1 capacity on work that should never have reached a model.



What the agencies recommend

The recommendations follow from the diagnosis. Design workflows that provide sufficient context while avoiding unnecessary prompts, excessive outputs, and repeated processing. Send relevant excerpts rather than whole files. Use caching for content that is processed repeatedly. Right-size the model to the task: "Using the same high-cost model across all workflows is not necessary." Recognize when traditional tools, smaller or locally run models, or manual processes are more efficient than a large model. Establish usage baselines, set alerts as consumption approaches its limits, and feed AI token telemetry into a security information and event management platform, where an unexplained spike can be investigated like any other anomaly. And optimize data before submission, with preprocessing, optimization, or filtering tools that "can help eliminate redundant content, reduce token consumption, and improve overall efficiency," assessed against the organization's risk management, security, privacy, and data protection requirements before they are deployed.

That last recommendation is the subject of Part III. Empathetic processing is a deterministic, security-aware way to do it for security data, and Parts III and IV show what it does to the arithmetic.

Cybersecurity technologist.

"Sending every security event to a large language model is an expensive way to avoid deciding what matters. The engineering work starts before inference: remove duplication, preserve the evidence, and supply the context the decision requires. Efficiency earns its keep when we can demonstrate that the savings haven't created a blind spot."

TK Keanini, cybersecurity technologist and author of tkonsecurity.com

The agencies did not review this paper, and their guidance does not endorse WitFoo or any product. The references here are to the public guidance, which is marked TLP:CLEAR.

The verdict from the agencies

When the national cyber security agencies of five countries publish guidance on token consumption, it is no longer a pricing question. It is an operational-resilience question, and cyber defense is the workload they most want protected.

PART II

Why AI-first SOCs break at scale

AI is not the problem. The architecture is. This section examines the dominant “feed everything to the LLM” pattern, shows the three taxes it imposes on any organization that adopts it, and works through a realistic cost example that most CISOs will recognize.

6 · The feed-everything-to-AI pattern

The dominant architecture pattern in AI-enabled security today is straightforward to describe: collect all security data, ship it into a SIEM or data lake, and point one or more AI systems at the whole corpus. Variations include routing SIEM alerts into an agentic AI SOC, running LLM summarization over incident bundles, deploying AI copilots for analyst queries, and using generative models to draft reports and communications.

Under this pattern, the AI is treated as a universal cognitive layer applied to a fully-detailed signal stream. The logic is appealing: more context produces better reasoning, and LLMs are now good enough that you can hand them raw data and expect coherent analysis. And in a controlled evaluation — one SOC, one week, one vendor’s demo dataset — this pattern produces very impressive results.

What the demos do not show

Demos operate on curated data. The logs are clean, the alerts are high-signal, and the volumes are manageable. When the same architecture is deployed against an actual production data stream, three problems emerge that were never visible in the evaluation.

The alerts are not high-signal. Industry research consistently shows enterprise SOCs processing one hundred thousand alerts or more per day, with one to five percent true-positive rates. The LLM is being asked to reason about noise ninety-five percent of the time.

The data contains massive redundancy. Single security events frequently appear in three, five, or ten different log streams — the firewall sees the connection, the proxy sees the request, the EDR sees the process, the cloud provider sees the API call, the identity provider sees the authentication. Without deduplication, the LLM is being charged to reason about the same underlying event multiple times.

The data is structurally inconsistent. Every vendor’s log format is different. Timestamps, field names, event types, severity levels — none of it is normalized. A significant fraction of the AI’s effective work becomes translating between vendor schemas rather than actually analyzing threats.

> **The demo–production gap**
>
> The pattern that looks brilliant on ten thousand curated alerts does not scale linearly to ten million uncurated ones. It scales super-linearly — because the noise, redundancy, and inconsistency compound as volume grows.

7 · The three taxes — redundancy, noise, retrieval

Every token of input or output sent through an LLM has a direct dollar cost. When security data is fed to an AI layer without pre-processing, three distinct taxes apply to that cost. Understanding them individually is essential to seeing why empathetic processing produces such dramatic economic leverage.

The redundancy tax

Security data is extraordinarily redundant. A single user logging into a single cloud application generates events in identity logs, proxy logs, cloud audit logs, DLP logs, EDR logs, and often a SIEM-internal correlation log. The underlying event is one login. The raw data representation is six to ten log entries. When that data flows into an AI layer without deduplication, the model is charged to reason — in input tokens and, often, in output tokens — about the same underlying event six to ten times.

At enterprise scale, the multiplier is even worse. A medium-sized organization generating five hundred gigabytes of security telemetry per day typically carries a high raw-to-unique signal ratio: the same event arrives many times over. Without deduplication, the LLM pays for every one of those repeats.

> "As a cybersecurity specialist organisation operating through 40 years of technology cycles, we have learned to distinguish architectural shifts from pricing ones. Pricing moves in both directions; architecture, once it is right, compounds. The question we have been working through is not whether AI belongs in security operations, but where in the pipeline it earns its cost, and what has to be true of the data before it gets there. Cybersecurity data is structurally redundant. The same underlying event surfaces in half a dozen log streams, and routing that raw stream into expensive compute, whether an LLM or a SIEM, is a unit-economics problem regardless of where token prices eventually settle. Empathetic processing addresses the problem at the correct layer: deterministic, security-aware curation before inference. In our pipelines it has delivered data-rate reductions above 90% with no loss of analytic value, and the artifacts our analysts and downstream models work from are materially richer. That is a structural improvement in how we can scale on behalf of our clients."
>
> **Luke Taylor**, Chief Executive Officer, Scientific Software and Systems

The noise tax

Even after deduplication, most security data is noise. Research across many enterprise SOCs consistently shows that one to five percent of alerts are true positives. The other ninety-five to ninety-nine percent are false positives, benign anomalies, or low-value telemetry. An AI SOC platform that analyzes every alert is spending the overwhelming majority of its compute — and its bill — on events that do not warrant analysis.

Agentic workflows make this tax worse, not better. An agent investigating a false-positive alert typically makes nearly as many LLM calls as one investigating a real incident, because the agent has to run the investigation before it can determine the alert is benign. The false positive may require slightly shorter reasoning traces, but it still costs real money. Ninety-five percent of an AI SOC's token budget goes to proving that nothing was wrong.

The retrieval tax

Retrieval-augmented generation (RAG) has become the dominant architecture for giving LLMs access to security context. The model is paired with a vector database of historical events, policies, threat intel, and configuration state. For each query, the system retrieves relevant chunks and includes them in the context window.

This is effective — but expensive. The retrieval step is cheap; the subsequent inclusion of retrieved chunks in every LLM call is not. A typical security investigation may pull in five to twenty context chunks averaging five hundred to two thousand tokens each, multiplying input tokens by five to forty per round. With agents that iterate across multiple steps, each step reloading context, the retrieval tax compounds. Context windows in excess of 200 thousand tokens — which Gemini 2.5 Pro and several competitors support — compound it further, because retrieval naturally fills the available space.

A fourth exposure: an attacker's hand on the meter

The agencies' guidance names a risk the three taxes do not capture. In a denial-of-wallet attack, threat actors exploit consumption-based pricing to drive up an organization's costs; in a denial-of-tokens attack, they generate enough usage to exhaust its allocations or trigger its rate limits. Security operations are unusually exposed to both, because an adversary controls part of a SOC's input. Every failed login, port scan, and noisy payload an attacker sends becomes telemetry, and in a feed-everything architecture every alert it triggers becomes a prompt, an investigation, and a line on the token bill.

Alert flooding has always been a way to hide an intrusion in noise. Under per-token pricing it is also a way to drain a budget, or exhaust a quota, at the moment the SOC most needs capacity. The exposure scales with whatever the attacker can generate, not with what the defender can afford.

A deterministic layer in front of the model changes that arithmetic. Repeated reports of the same interaction update one artifact's counts and timestamps instead of producing new prompts. An attacker who varies every event can still create new artifacts, so this is not immunity. But classification and incident structuring decide how many artifacts reach expensive reasoning, and the token spend that remains tracks distinct activity rather than raw event volume.

What the three taxes look like combined

These three taxes interact, and their combination is the right way to reason about the total cost gap between a raw-data AI architecture and a curated one. A caveat matters first: the taxes are correlated, not independent. Duplicated events are disproportionately low-value noise; retrieval cost scales more closely with the number of events actually investigated than with raw ingest volume. A naive multiplication of 10× redundancy by 20× noise by 10× retrieval would produce a 2,000× theoretical upper bound, but this overstates the real-world effect because the underlying distributions overlap substantially.

A more defensible characterization: when the three taxes are modeled with realistic correlation — using WitFoo production deployment data and third-party studies of SIEM redundancy, alert quality, and RAG token overhead — the combined multiplier falls in the range of 50× to 300× for typical enterprise workloads, with higher figures observed in environments with extreme alert noise or heavy agentic use. Even at the low end of that range, the economic gap is transformative; the paper's conclusions do not depend on the upper end.

50 – 300×

Typical combined cost multiplier of redundancy, noise, and retrieval over a curated, pre-AI baseline — modeling observed correlation between the three taxes

Put differently: an AI SOC architecture that applies an LLM to raw, unfiltered, unnormalized security data is typically paying one to two orders of magnitude more than it would pay to produce the same analytical outcomes against a clean, deduplicated, normalized signal stream. That is the economic gap empathetic processing closes — and it is more than enough to carry the rest of the argument.

8 · A worked cost example

To make the economics concrete, consider a mid-sized enterprise with roughly five thousand employees. The security posture is typical: standard endpoint and network tooling, cloud workloads in two or three providers, a Splunk-class SIEM, and a recently adopted AI SOC platform for Tier-1 triage. Daily raw security data volume is five hundred gigabytes.

Baseline: unfiltered AI-first architecture

In this architecture, security data flows into the SIEM. Alerts generated by detection rules flow to the AI SOC platform, which performs agentic investigation. The AI SOC uses a mid-tier frontier model (Sonnet-class pricing: $3 per million input tokens, $15 per million output) for routine triage and a premium frontier model (Opus-class or GPT-5.2: average $10 per million input, $50 per million output) for escalated investigations.

Cost component	Assumption	Monthly cost (USD)
SIEM ingest licensing	500 GB/day @ typical $8/GB/day	$120,000
SIEM hot storage (30-day retention)	15 TB hot tier	$22,000
SIEM cold storage (13-month)	190 TB cold / archive	$18,000
AI SOC platform subscription	Enterprise tier	$25,000
AI inference — mid-tier (triage)	100K alerts/day × 15K tokens avg	$38,000
AI inference — frontier (investigation)	2K escalations/day × 40K tokens avg	$42,000
Vector database + retrieval infra	Managed RAG stack	$8,000
Monthly total		**$273,000**
Annualized		**$3.28M**

Figures are illustrative, based on published SIEM ingest pricing, representative AI SOC platform pricing, and January–April 2026 LLM API rates. Actual enterprise numbers vary with vendor, region, and negotiated discounts.

The triage line assumes that the repeated context in each triage prompt (more than four-fifths of it) is served from the provider's prompt cache at a 90 percent discount, the practice the agencies' guidance recommends. Without caching, the same line would run to more than $130,000 a month.

The cost breakdown that matters

In this architecture, AI inference — the two line items for mid-tier triage and frontier investigation — accounts for roughly thirty percent of the monthly bill and is the fastest-growing component. As data volumes expand thirty to forty percent per year and frontier token prices hold steady, the AI inference line will grow faster than the SIEM line. Within three years, at these assumptions, the AI portion alone would approach two million dollars per year.

Worse: most of that AI spend is being paid to analyze noise. At a one-percent true-positive rate across one hundred thousand alerts per day, the AI is producing roughly one thousand real investigations per day. The other ninety-nine thousand alerts consume input, retrieval, and output tokens only to conclude that nothing was wrong. The marginal cost of reaching that conclusion is approximately eighty percent of the AI inference bill.

What scale makes worse

Now extrapolate. If this organization grows to ten thousand employees, or adds three more business units, or migrates a data center to cloud — any of the standard growth vectors that drive security data upward — volumes can easily double. At double the raw data, most of the line items in the baseline table grow linearly or super-linearly. The ingest cost doubles. The storage cost doubles. The AI triage cost grows faster than doubles because alert rates expand non-linearly with data volume.

At that trajectory, the security program is looking at seven million dollars per year, growing twenty to thirty percent year over year. This is unsustainable for most enterprises. And it is entirely the result of an architecture that routes unfiltered data to expensive compute.

Security leadership in higher education.

“Universities run some of the most open and varied networks anywhere, on security budgets that never grow as fast as the logs do. If every event goes to an AI model, the token bill becomes one more cost competing with teaching and research. Shaping the data first, while keeping every record traceable, is how a campus security team gets AI’s help without trading away evidence or budget.”

Leilani Lauger, Higher Education CISO

The structural problem

It is not that AI is too expensive. It is that AI is being asked to do work that a security-aware, deterministic layer could do for a thousandth of the cost. The question is not how to negotiate better AI pricing. The question is how to make AI do dramatically less work while producing the same or better analyst outcomes.

PART III

Empathetic processing

The answer to the cost collision is a different kind of processing layer — one that understands security data semantically, deduplicates based on what events actually mean, and delivers AI a dramatically smaller, richer signal to reason about. This section defines it, explains how it works, and shows where AI still earns its keep.

9 · What empathetic processing means

Conventional data pipelines for security — the layer that sits between your sources and your analysis — are indifferent to what the data means. They accept log lines, parse them into fields, forward them onward. Routing decisions are based on tags, patterns, or rules written by the operator. This is what Cribl, Splunk's Ingest Actions, Microsoft's Data Collection Rules, and most log collection frameworks do.

Empathetic processing is different. It is a data layer that understands, at the moment of ingestion, what security evidence looks like. It knows that an authentication event has a user, a source, a destination, and a timestamp. It knows that a firewall connection involves a source, a destination, a port, and often a user or a service. It knows that the same underlying real-world event is often represented six different ways across six different log streams. And it knows which representations matter for security and which do not.

The word "empathetic" matters

We chose the word deliberately. An empathetic system is one that understands the meaning and context of what it is processing, not just the syntax. It recognizes intent, not just pattern. It knows what a SOC analyst would do with this evidence — and prepares the evidence accordingly.

Applied to security data, empathy means three things: preserving every piece of evidence that a real investigation might need, collapsing every piece of data that does not add evidentiary value, and normalizing every representation so downstream systems see consistent structure. The goal is not less data. The goal is less noise, without losing a single fingerprint.

Why this matters for AI economics

AI systems — whether a frontier LLM, an embedding-based classifier, or an agentic workflow — are most expensive when they have to do semantic work that deterministic logic could do faster and cheaper. Parsing a Cisco ASA log format, detecting that two log lines describe the same TCP session, recognizing that a DNS query and a TCP connection one millisecond later belong to the same transaction: none of this requires an LLM. It requires a well-designed, security-aware processing layer.

By moving this work out of the AI and into deterministic processing, empathetic layers free the AI to do what it is uniquely good at: understanding attacker intent in messy human-generated contexts, correlating evidence across disparate systems, producing narratives that humans can act on. The AI budget goes up in value per dollar by orders of magnitude because the AI is no longer doing cheap work at expensive prices.

> **The division of labor**
>
> Deterministic processing is cheap, predictable, and fast. AI reasoning is expensive, flexible, and adaptive. An empathetic pipeline puts each where it belongs: deterministic logic handles the 95–99% of work that is pattern-bound, and AI handles the 1–5% that genuinely requires open-ended reasoning.

10 · ProtoGraph deduplication

The core deduplication technique inside WitFoo Conductor is called ProtoGraph. It is the primary mechanism by which Conductor makes what reaches the systems downstream more than 90 percent smaller, the reduction that transforms the downstream economics. That is not sampling. Nothing security-relevant is discarded.

What ProtoGraph does

ProtoGraph maintains a lightweight, continuously-updated graph of security-relevant entities and the relationships between them. The nodes are the things security cares about: users, hosts, processes, files, network endpoints, cloud resources, identities. The edges are the observed interactions: logins, connections, executions, access grants, authentications.

When a new log event arrives, Conductor parses it, identifies the entities involved, and checks the graph. If the event describes a relationship already captured — the same user logging in from the same host, the same connection between the same source and destination, the same repeated telemetry heartbeat — the event is collapsed into the existing graph structure. The evidence is preserved, counts and timestamps are updated, but no new record is emitted. If the event describes something new — a previously-unseen entity, a novel relationship, a changed attribute — a new structured artifact is emitted downstream.

The result is that hundreds of log events describing a single underlying security-relevant fact become a single, richly-structured artifact, with all contributing sources accounted for. Downstream systems — whether SIEM, data lake, AI SOC platform, or WitFoo Analytics — see a dramatically smaller, dramatically cleaner signal stream.

Why security-aware dedup outperforms generic dedup

Generic observability pipelines perform deduplication based on byte-level or hash-level matching. Two identical log lines collapse into one. This is useful for reducing accidental duplicates (the same log forwarder configured twice, for instance), but it does nothing about the structural redundancy that dominates security data.

Consider a single user logging into a single SaaS application. The identity provider emits an authentication event with one schema. The cloud proxy emits a traffic event with another. The SaaS platform emits an API event with a third. The endpoint emits a process event with a fourth. None of these log lines are byte-identical. A generic deduplicator cannot collapse them. A security-aware deduplicator recognizes that they describe the same underlying interaction and represents them as one enriched artifact with four source attributions.

This is the structural win. The same deduplication mechanism that reduces volume also enriches the remaining artifacts with cross-source context — exactly the context an AI system or analyst would otherwise have to assemble itself, often by making additional expensive queries.

What ProtoGraph preserves

A frequent concern is whether aggressive deduplication discards forensic evidence. ProtoGraph is designed to answer this concern emphatically: every contributing event's source, timestamp, and distinguishing attributes are retained in the graph structure. An investigator can drill into the original per-source details at any point. Compliance auditors can reconstruct the full raw signal if required. The difference is that during normal operation — triage, investigation, reporting — downstream systems never have to wade through the redundancy to find the signal. The deterministic entity-identity and graph-merge machinery underneath ProtoGraph — the mechanism that guarantees the same real-world entity resolves to the same graph identity without loss of evidence — is the subject of a U.S. provisional patent application filed with the USPTO in July 2026 (patent pending).

> **The compound effect**
>
> Because ProtoGraph both reduces volume (more than 90%) and enriches context (cross-source attribution, entity relationships, normalized structure), it attacks all three of the taxes from Chapter 7 simultaneously. Redundancy goes down by an order of magnitude. Noise goes down because low-value repetitive events are collapsed. Retrieval cost goes down because each artifact already carries the context that RAG would otherwise have to fetch.

11 · The pre-AI curation discipline

ProtoGraph deduplication is the volumetric workhorse, but empathetic processing involves several additional disciplines that compound its impact. Together, these transform a raw security data stream into something that AI can reason about efficiently — and that a deterministic system can often handle entirely without AI.

Adaptive parsing

Every security data source has its own log format. Firewalls from Palo Alto, Fortinet, Check Point, and Cisco each produce different field layouts. Cloud providers emit JSON structures that change between service versions. Endpoint agents produce structured and semi-structured output in competing conventions. Traditional SIEM deployments handle this with hand-written parsers, which require ongoing maintenance as vendors change formats.

Conductor uses an NLP-based adaptive parsing approach. New log formats are classified automatically: the system recognizes the kind of event (authentication, network, process, file, identity) and extracts the relevant fields without requiring a pre-written parser. When an entirely novel format appears, WitFoo can deliver updated parsing in approximately three business days. The operational effect is that the customer never maintains a parser.

For AI economics, adaptive parsing matters for a reason that is not immediately obvious. The most expensive thing you can do with an LLM is ask it to parse structured data from unstructured text. Every token spent getting the model to extract an IP address or a username from a log line is a token that could have been spent on actual analysis. Adaptive parsing removes this cost entirely.

Normalization

Once parsed, events are normalized into a common schema. Source, destination, user, process, file, outcome, timestamp — regardless of which vendor's product emitted the event, the downstream consumer sees the same fields in the same places. This is the single largest operational gift Conductor gives to any downstream AI system.

A detection rule written against the normalized schema works across every vendor. A prompt written against normalized events produces consistent results regardless of source. An agentic workflow investigating a login event does not need to know whether the event came from Okta, Duo, Azure AD, or a custom SSO — the shape is identical.

Enrichment

Conductor attaches contextual enrichment to every emitted artifact: threat intelligence tags, geolocation, asset classification, user role, known-good or known-bad indicators. This enrichment is performed once, at ingestion, and attached to the artifact. Downstream systems — and AI in particular — no longer need to perform separate retrieval calls to fetch this context.

This is the direct counter to the retrieval tax discussed in Chapter 7. If a security-relevant artifact already carries the contextual attributes an analyst (or an agent) would look up, the retrieval step is unnecessary. Input token usage per analytical operation drops sharply.

Classification

The final discipline is lightweight classification. Conductor attaches a signal quality and category label to every artifact: this is a routine authentication (green), this is a notable anomaly (yellow), this appears to be a security-relevant event worth investigation (red). Downstream AI systems — including WitFoo Analytics and any third-party AI SOC platform — can use these labels to route work intelligently. Only the red-labeled artifacts need frontier-model reasoning. Yellow can often be handled by mid-tier models or deterministic rules. Green is logged for compliance but rarely requires AI attention.

~95%

of analytical work can be resolved with curated, normalized, enriched signals — without ever reaching frontier inference

What the agencies call optimizing data before submission

The agencies' guidance describes this layer in general terms. It suggests that organizations use preprocessing, optimization, or filtering tools to reduce unnecessary data before it is submitted to an AI system, tools that "can help eliminate redundant content, reduce token consumption, and improve overall efficiency." It also says any such tool should be assessed against the organization's risk management frameworks and its security, privacy, and data protection requirements before deployment. That is the right test, and a deterministic layer lends itself to it: its behavior is repeatable, and every record it emits points back to the raw evidence it came from.

> *Chief information security officer with thirty-four years in the security industry.*
>
> *"Charles Herring solved the right problem before most of the industry understood what the problem was. The AI cost collision in cybersecurity is not a pricing negotiation — it is a structural indictment of architectures that route unfiltered noise to expensive reasoning engines and call it intelligence. Empathetic processing is the clinical answer to a clinical problem: you cannot treat a patient accurately when the instruments measuring them are generating 95% false signal. WitFoo Conductor is not a data pipeline. It is the pre-AI curation discipline that determines whether your security dollar buys prevention value or buys the illusion of it.*
>
> *"I have spent thirty-four years watching the security industry mistake activity for work, and volume for intelligence. Charles has built the instrument that makes the distinction measurable. In a post-Mythos world where the attacker's execution chain runs faster than most defenders can detect, the organisation that feeds clean, deduplicated, velocity-aware signal to its AI layer will survive the transition. The organisation feeding a firehose will not, not because the AI failed, but because you cannot reason your way to good outcomes from bad inputs.*
>
> *"AI cannot solve an epistemological problem with a computational solution. Garbage in, intelligent certainty out. Charles understood this. He built accordingly."*
>
> **Gopal Padinjaruveetil**, KnowNothingCISO

12 · Temporal Link Analysis — from signals to structured incidents

ProtoGraph deduplication and the curation discipline together produce a stream of clean, normalized, enriched artifacts. That is a dramatic improvement over raw logs. But an artifact is still not an incident. An investigator does not reason about individual artifacts; they reason about coherent stories — sequences of related events that fit together into an attack narrative, a policy violation, or a business-as-usual pattern. The final discipline in the empathetic processing stack is the one that turns artifacts into structured incidents. WitFoo calls this Temporal Link Analysis.

The four-layer incident structure

Temporal Link Analysis (TLA) continuously evaluates the artifact stream and, when a cluster of related activity is detected, constructs a four-layer representation of it. Every incident produced by TLA carries all four layers.

- **Graph nodes —** the entities involved in the incident: users, hosts, processes, files, network endpoints, cloud identities, SaaS resources. Each node carries its enrichment (asset classification, user role, known-good or known-bad reputation) and a stable identifier that ties it to the same entity wherever it appears elsewhere in the graph. The method by which TLA composes an incident's risk score from its underlying evidence, weighing how completely each candidate theory covers known attack patterns, is also the subject of a July 2026 U.S. provisional patent application (patent pending).
- **Graph edges —** the observed interactions between entities, ordered in time and, where determinable, causally linked. An edge records what happened (login, connection, execution, access, authentication), when it happened, which artifact observed it, and what followed from it.
- **Theories of crime —** candidate explanations for the observed pattern, scored against reference models. TLA evaluates each incident against MITRE ATT&CK technique chains, known threat-actor playbooks, policy violation templates, and benign-but-unusual baselines. Each theory is ranked by how well the graph's evidence supports it.
- **Underlying artifacts —** every source event that contributed to the graph, with attribution preserved. An investigator, an auditor, or an AI system can drill from any node, edge, or theory back to the specific raw evidence that produced it. No claim in the incident is unsupported; no artifact is orphaned from the conclusions it informs.

Why structure transforms AI quality

A structured TLA incident does not just compress the data an AI system has to read. It transforms the kind of work the AI is asked to do — from discovery to evaluation.

Without TLA, an AI tasked with understanding a multi-source security event has to perform the discovery work itself: read through thousands of log lines, identify the entities, infer the relationships, hypothesize the attack pattern, and then reason about what it all means. Every step in that chain is a place where the model can hallucinate, miss context, or draw a subtly wrong conclusion.

With TLA, the discovery is already done, deterministically. The entities are identified. The edges are ordered. Candidate theories are enumerated. The AI's job becomes evaluation: given this structured graph, which theory best fits? Is anything missing? Produce a SOC-ready narrative for the most likely theory. This is a bounded, verifiable prompt — the kind of work LLMs do exceptionally well — rather than an open-ended reasoning task that stresses every limit of the model.

Defensibility and auditability

The most underappreciated property of TLA is that it makes AI-assisted incident response defensible. Because every claim in the incident graph traces to specific artifacts, and every AI output is generated from that graph, the full reasoning chain is auditable. A regulator, a compliance officer, or a legal team can follow the thread from "the AI concluded X" back through "the graph encoded Y relationship" to "these specific events supply the evidence."

In regulated industries — finance, healthcare, defense, critical infrastructure — this property is not a nice-to-have. Regulators are beginning to require explainability and provenance for AI-assisted decisions in security operations. A SOC architecture that cannot produce an evidence trail for every AI-generated conclusion will fail the next generation of audits. TLA is the layer that keeps the AI output anchored to ground truth.

Why structure collapses token cost

The token economics of TLA are as compelling as the quality economics. A moderately complex incident — the kind that draws from ten or more data sources across a thirty-six-hour window — can easily involve forty thousand or more raw log entries. Sent to an LLM as a flat context, that is hundreds of thousands of input tokens, a generation task running for thousands of output tokens, and (often) multiple agentic rounds of retrieval and re-prompting to get the model to converge on a useful answer.

A TLA representation of the same incident is a structured object: a few hundred nodes, perhaps a thousand edges, two to five candidate theories with evidence pointers, and a compact metadata header. Serialized for an LLM, it fits comfortably in five to fifteen thousand tokens. The prompt that accompanies it is bounded: evaluate the theories and produce a narrative. One clean round of inference. One output generation.

Scenario	Input tokens (typical)	Output tokens (typical)	Inference rounds
Raw evidence to LLM (no TLA)	150,000 – 400,000+	2,000 – 8,000	3 – 10
TLA-structured incident to LLM	5,000 – 15,000	800 – 2,000	1 – 2
Effective reduction	**~20 – 30×**	**~3 – 5×**	**~3 – 5×**

Ranges reflect observed variation across WitFoo deployments. Compounded across all three dimensions (input, output, and round count), the per-incident token cost reduction commonly falls in the 30 – 100× range for complex incidents — and the quality of the output improves at the same time.

The prompt transformation

The easiest way to see what TLA changes is to compare the prompt an LLM receives in each architecture. The shape of the prompt reveals the shape of the work.

> **Without TLA**
>
> *Here are 42,867 log entries across 11 data sources covering the last 36 hours of activity on our network. Identify any potential security incidents, explain what happened, score the likelihood of compromise, and produce a narrative suitable for the SOC manager. Use your judgment about which events are related.*

> **With TLA**
>
> *Incident candidate: 340 nodes, 1,120 edges, 4 theories of crime scored against MITRE ATT&CK. All supporting artifacts are attached as evidence pointers. Which theory best fits the evidence? Identify any gaps. Produce a SOC-ready narrative for the leading theory.*

One is an expensive, unbounded, error-prone prompt that leans on the LLM to do work it is poorly suited for. The other is a cheap, bounded, verifiable prompt that plays to the LLM's strengths. The economic gap between them is the same gap that separates an AI-enabled SOC that renews painfully from one that renews comfortably.

30 – 100×

Typical per-incident token cost reduction from routing AI through TLA-structured incidents rather than raw evidence

How TLA fits the empathetic processing story

The three disciplines of this Part compound. ProtoGraph removes redundancy at the signal level. The curation stack normalizes, enriches, and classifies what remains. Temporal Link Analysis synthesizes the curated signals into incident-shaped structures that AI can reason about cheaply, accurately, and defensibly. Each layer makes the next one more effective. Together they convert a data problem that AI cannot afford to solve into one that AI is ideally positioned to solve.

> **The division of labor, revisited**
>
> Deterministic logic does the discovery: identifying entities, ordering events, enumerating theories, preserving evidence. AI does the evaluation: judging which theory fits, filling narrative gaps, communicating findings. This division is not just economically efficient. It is epistemically correct. Deterministic systems are better at structured extraction; AI systems are better at language and judgment. Empathetic processing simply lets each do what it does best.

13 · Where AI earns its cost

Empathetic processing is not an argument against AI. It is an argument for deploying AI where AI produces unique value — and nowhere else. This chapter lays out the four classes of security work where AI reasoning is genuinely irreplaceable, and why routing those specific workloads through AI produces exceptional returns once the rest of the pipeline has been optimized.

The agencies' guidance reaches the same division of labor from the other side. "Using the same high-cost model across all workflows is not necessary," it says; organizations should right-size the model to the task, set the reasoning or effort level to limit costly output, and recognize when traditional tools, smaller or locally run models, or manual processes are more efficient than a large model. For cyber defenders it recommends a multi-model approach in which smaller models complete less complex security tasks and larger models and agents are reserved for complex ones. Classification at the empathetic layer is what makes that routing practical in a SOC: routine artifacts rarely reach a model, notable ones can go to a smaller one, and only security-relevant ones earn frontier reasoning.

Investigation depth

The single most valuable thing an AI system does in a security operation is expand the depth of investigation available per incident. A well-designed agentic workflow can pull in threat intel, historical context, asset information, user history, network topology, and recent related alerts — and synthesize all of it into a coherent narrative — in seconds. This is work that a human analyst would take thirty to ninety minutes to do manually.

For this use case, AI cost is justified and the return is enormous. But the economics only work when the AI is investigating real incidents, not false positives. An empathetic pipeline that surfaces roughly five hundred to one thousand genuinely-notable artifacts per day — down from one hundred thousand raw alerts — turns this use case from economically unsustainable to economically compelling.

Cross-source correlation

The second high-value AI use case is correlation across sources that do not share a common identifier. A login from an unusual location, a subsequent data access on a sensitive asset, and a later outbound connection to a previously-unseen external endpoint — each of these might individually be benign, but the sequence across three different systems tells a different story. Traditional correlation engines can catch sequences that match pre-written rules. AI can catch sequences that no rule has ever been written for.

This only works if the AI sees normalized, enriched signals. If the AI is spending effort to parse vendor-specific fields and look up missing context, it never gets to the actual correlation. Empathetic processing removes that overhead.

Natural-language interaction

Analyst productivity gains from AI are real and substantial when the AI can answer natural-language questions about the security state. “Show me every user who accessed the payroll system from a new device this week.” “What is the blast radius if this endpoint is compromised?” “Has this TTP pattern appeared before in our environment?” These queries are where AI SOC copilots shine.

Again, the quality of the answer depends on the quality of the underlying data. A copilot querying a deduplicated, normalized, enriched artifact store produces accurate answers in one or two round trips. The same copilot querying raw logs requires many more round trips — each with higher input token counts — and the answers are less reliable. Empathetic processing makes copilot-style interaction both cheaper and better.

Report generation

Finally, AI is uniquely good at producing human-readable output: incident narratives, executive summaries, compliance reports, customer communications. This is genuinely time-saving work — turning a structured incident record into a paragraph that a board can read, or a regulator can consume, takes analysts real time, and AI does it in seconds.

Output tokens are the expensive part of generation work — 3 to 8 times the cost of input. Empathetic processing does not reduce output cost; it reduces input cost and the number of generation rounds needed. While a narrative generated from a raw alert bundle requires many rounds of clarification, correction, and re-prompting, a narrative generated from a dense, pre-enriched artifact requires one clean round of inference.

The AI-at-its-best pattern

The economic sweet spot for AI in security operations: deep investigation on a curated set of real incidents, cross-source correlation on normalized signal, natural-language interaction with enriched artifacts, and polished output generation from dense inputs.

Empathetic processing is what makes this sweet spot reachable. Without it, AI cost expands to fill the available budget. With it, AI cost is contained to the small fraction of work where AI genuinely outperforms deterministic logic — and the quality of that work is better because the inputs are cleaner.

PART IV

The economic math

Everything above this section is argument. This section is numbers. We lay out the volume savings, the token savings, the storage and compute savings, and the pricing model that shields customers from every one of the cost pressures Part I described.

14 · Data volume reduction

The single most important number in WitFoo economics is the post-Conductor volume reduction. What reaches the systems downstream is more than ninety percent smaller, and every record in it points back to its raw evidence. That is not sampling. Nothing security-relevant is discarded. The specific figure for any given customer depends on source mix, environmental characteristics, and how much redundancy already exists in the raw data stream. For the examples in this section, we use ninety percent, the floor of the figure WitFoo publishes.

Translating percentage into dollars

Applied to a baseline ingest volume, a ninety percent reduction transforms the downstream cost structure completely. The example organization from Chapter 8 — five hundred gigabytes per day — becomes fifty gigabytes per day of structured, deduplicated, enriched artifacts flowing to the SIEM, the AI SOC platform, or both.

Metric	Raw (pre-Conductor)	Post-Conductor	Reduction
Daily volume	500 GB	50 GB	**90%**
Daily events (typical ratio)	~500 million	~50 million	**90%**
Alerts requiring model triage	~100,000	~5,000	**95%**
Hot storage (30-day)	15 TB	1.5 TB	**90%**
Cold storage (13-month)	190 TB	19 TB	**90%**

The alert reduction is the most important number

Notice that the alert count drops further than the volume does: roughly ninety-five percent of today's alerts never need model triage. This is because Conductor is not just collapsing redundant log entries; it is correlating events into semantically meaningful artifacts. What appeared as one hundred thousand separate alerts is revealed as a few thousand artifacts worth a model's attention, and perhaps one thousand real security-relevant incidents, each one richer and better-contextualized than any single alert in the raw stream.

This number drives the AI economics. An AI SOC platform charged to investigate one thousand real artifacts per day is performing useful work. An AI SOC platform charged to investigate one hundred thousand raw alerts, ninety-nine percent of which are false positives, duplicates, or low-value telemetry, is wasting ninety-nine percent of its budget proving that nothing was wrong.

15 · Token savings

Data volume reduction translates into token savings through several distinct mechanisms. Each one compounds. Understanding them individually is essential for modeling the economic impact of empathetic processing on any specific workload.

Direct input token reduction

The first-order saving is obvious: if a workflow is ingesting five hundred gigabytes of context per day and Conductor cuts that to fifty gigabytes, the input tokens feeding any subsequent AI analysis drop by ninety percent. At typical mid-tier pricing ($3 per million input tokens) and enterprise scale, this alone is worth tens of thousands of dollars per month.

Retrieval cost reduction

The second-order saving comes from enrichment. Because Conductor attaches threat intel, asset context, user role, and cross-source attribution to every emitted artifact, downstream AI systems need to perform fewer retrieval operations. Each retrieval step that would have pulled five to twenty context chunks into the input window — multiplying input tokens by that factor — is now unnecessary. This reduces input token consumption per analytical operation by another large factor.

Generation round reduction

The third-order saving is in output tokens. Cleaner, denser input produces better first-pass output from any LLM. Agentic workflows that would have iterated through three, five, or ten rounds of investigation, clarification, and correction frequently resolve in one or two rounds when the input artifact is well-structured. Because output tokens cost three to eight times more than input tokens, reducing output generation rounds has disproportionate impact on the total bill.

Worked token math

Apply these three mechanisms to the Chapter 8 baseline. The baseline AI inference line items (triage plus investigation) were approximately $80,000 per month. With empathetic processing in front:

Component	Baseline monthly	With empathetic processing	Reduction
AI — triage (mid-tier)	$38,000	$1,800	**~95%**
AI — investigation (frontier)	$42,000	$9,500	**~77%**
Vector DB / retrieval infra	$8,000	$2,000	**~75%**
AI subtotal	**$88,000**	**$13,300**	**~85%**

The triage line drops sharply because the volume of alerts needing triage is reduced by roughly ninety-five percent. The investigation line drops less sharply because the number of real investigations does not drop proportionally — a serious incident is still a serious incident — but the quality of the input makes each investigation cheaper to run. The retrieval line drops because enrichment replaces most external context lookups.

$74,700

Monthly AI cost reduction in the worked example (~85% savings) — before SIEM savings, before storage savings, before headcount effects

> "When log volume drops by more than 90 percent and each remaining event comes enriched with user, asset and threat intel context, an AI agent can reach a conclusion in a handful of calls instead of dozens. That's where the token savings come from. Additionally, storage needs and costs drop significantly while data quality and compliance stay intact."
>
> **Thomas Whang**, Chief Technology Officer, Impelix

16 · Storage, compute, and transit savings

The AI savings alone would justify empathetic processing in most enterprise contexts. But they are not the whole story. Data volume reduction cascades through the rest of the security infrastructure stack with similar impact.

SIEM ingest and storage

Splunk, and most of its competitors, charge by daily ingest volume. A ninety percent reduction in volume flows directly to a ninety percent reduction in SIEM ingest licensing cost — often the single largest line item in any enterprise security budget. Storage costs (both hot tier for active search and cold tier for retention) drop in proportion. Indexer count, search head count, and compute required to run the SIEM drop alongside.

For the Chapter 8 baseline organization, the SIEM line items ($160,000 per month combined) fall to approximately $16,000 per month after Conductor. Whether that saving is realized depends on the customer's existing contract structure — some SIEMs have minimum commits — but the trajectory over any renewal cycle is dramatic.

Network and transit

Multi-site enterprises often discover that the WAN bandwidth dedicated to security log forwarding is a meaningful fraction of total bandwidth cost. Conductor's ability to process data locally at each site — reducing volume before the WAN link — cuts this bandwidth proportionally. Branch offices on metered links, remote sites with satellite connectivity, cloud workloads with egress-based pricing: every one of these see direct savings.

> *Cybersecurity engineering for federal and defense customers.*
>
> *"For eight years Ardalyst has worked alongside WitFoo on empathetic processing and CyberGrid workflows across the constrained, unreliable networks our federal and defense customers operate on every day. The research has unlocked workflows that were simply not possible with legacy tooling — and has done so at a fraction of the compute footprint, with almost no ongoing maintenance burden on our engineers."*
>
> **Josh O'Sullivan**, Chief Technology Officer, Ardalyst

Compute infrastructure

Self-managed SIEM clusters, data lake query engines, and analytics platforms all size compute capacity to ingest rate. When ingest drops, those clusters can be scaled down. Cloud-hosted SIEMs pass the savings through automatically. On-premises deployments realize savings at the next hardware refresh.

Headcount effects (qualitative)

The most important non-financial effect of empathetic processing is on analyst time. Traditional SOCs spend large fractions of analyst hours on alert triage, log query correction, and data format wrangling. Normalized, enriched artifacts make every analyst hour more productive. While we cannot quantify this universally — headcount economics vary too much across organizations to model — the qualitative effect is consistently reported by customers: analysts spend more of their time on actual security work and less on infrastructure fighting.

Total cost of ownership reframed

Apply all of the above to the baseline. Starting from $273,000 per month ($3.28 million per year) for the unfiltered AI-first architecture, the same security outcomes with empathetic processing in front look like this:

Component	Baseline monthly	With Conductor (Pro)	Reduction
SIEM ingest licensing	$120,000	$12,000	−90%
SIEM hot storage	$22,000	$2,200	−90%
SIEM cold storage	$18,000	$1,800	−90%
AI SOC platform subscription	$25,000	$25,000	—
AI inference (mid + frontier)	$80,000	$11,300	−86%
Retrieval / vector DB	$8,000	$2,000	−75%
WitFoo Conductor Pro (amortized)	—	$3,333	—
Monthly total	**$273,000**	**$57,633**	**−79%**
Annualized	**$3.28M**	**$692K**	**−79%**

Conductor Pro at $40,000 per year = $3,333 per month. Figures illustrative; actual savings vary. Existing SIEM contracts may have minimum commits that delay full realization of savings until renewal.

$2.58M, −79%

Annual savings in the worked example: the same security outcomes for about a fifth of the baseline cost

17 · The WitFoo pricing model

The final economic lever — and the one that compounds over time — is the WitFoo pricing model itself. Across the 7th Generation product line, every appliance is priced per unit, with unlimited data rates within each appliance's capacity envelope. There are no data volume charges, no per-alert fees, no per-token pass-throughs. A Conductor Pro appliance costs forty thousand US dollars per year regardless of how much data it processes inside that envelope.

Why the pricing model matters for AI economics

Most of the cost pressure described in Parts I and II comes from pricing models that scale with volume, activity, or token count. SIEM licenses scale with GB/day. AI SOC platforms scale with alert count. LLM APIs scale with tokens. Observability pipelines like Cribl scale with daily ingest. Storage scales with bytes retained.

In every one of these relationships, the customer's cost trajectory is tied to their data growth. If data grows thirty percent year over year — which most enterprise security programs are seeing — every one of those line items grows thirty percent year over year, even if the security outcomes do not.

WitFoo is the exception. Within each appliance's capacity envelope, a Conductor Pro absorbs volume growth at no incremental cost, and an Analytics Max handles query load at no incremental cost. Very high or rapidly growing volumes may eventually call for additional appliances, but appliance count scales in step-functions

that are far slower than linear with data volume. When the next big cloud migration or data center consolidation doubles the customer's security data volume, the WitFoo bill may move by a fraction of that — or not at all, if the existing appliances have headroom.

7th Generation pricing reference

The full 7th Generation product line and pricing is summarized below. Every listed price is per appliance, per annum, with unlimited data rates. Government, non-profit, and education customers receive a 15% discount on all published pricing, and additional appliances of the same product line in one order take volume discounts (Appendix A).

Conductor — data ingestion and signal processing

Tier	USD / year	NZD / year	Key features
Lite	$15,000	$19,500	Core integrations; ProtoGraph dedup; unlimited rates
Pro	$40,000	$52,000	All 30+ API integrations; STIX enrichment
Max	$60,000	$78,000	Pro features plus custom integrations

Reporter — compliance, tool effectiveness, cost analysis

Tier	USD / year	NZD / year	Key features
Lite	$20,000	$26,000	Bundled Conductor Lite; CSC8 compliance
Pro	$50,000	$65,000	Bundled Conductor Pro; 17 frameworks; cost/savings reports

Analytics — the full security operations platform

Tier	USD / year	NZD / year	Key features
Lite	$25,000	$32,500	Bundled Conductor Lite; signal search; incident response
Pro	$75,000	$97,500	Bundled Conductor Pro; SOAR; AI summaries; 17 frameworks
Max	$120,000	$156,000	Everything in Pro; multi-tenancy; AI-enabled playbooks

Console, the centralized management plane, is available at no cost and is recommended once a deployment reaches three or more appliances. Larger orders take the published volume discounts: every appliance of a product line after the first is 30, 50, or 65 percent off list, depending on how many of that line are in the order (Appendix A).

The compounding advantage

Over a five-year horizon with modest data growth (30% compounding annually), a customer locked into volume-based pricing pays about 2.9× their year-one price by year five for the same security outcomes. A WitFoo customer's appliance count may tick up modestly over that horizon, but the per-appliance price is protected against volume-driven inflation. List prices are subject to adjustment at renewal; multi-year price locks are available as a negotiated term on larger deals. Either way, the compounding gap versus volume-based pricing is the largest economic effect in the entire analysis.

Pricing is strategy

WitFoo's per-appliance, unlimited-data pricing is not just a sales differentiator. It is an economic hedge against every structural trend described in Part I of this paper. As AI compute tightens and per-token pricing holds, as SIEM licensing inflates, as data volumes expand — WitFoo's costs stay flat. The protection compounds with every year.

PART V

The strategic outlook

The present-day economics are compelling on their own. What makes empathetic processing strategically decisive — not just tactically useful — is the direction the AI market is heading. This section lays out the supply and demand trajectory, explains why WitFoo's advantages compound, and closes with recommendations for security leaders planning 2026 through 2029.

18 · The AI supply and demand curve

To reason about a three-to-five-year strategic horizon, it is useful to look at the supply and demand curves in AI compute separately — and then together.

Demand side

AI inference demand is being driven by four compounding vectors. Enterprise adoption is still in early innings: most organizations have deployed AI in narrow pilot use cases and are now scaling to production workloads. Agentic architectures multiply the LLM calls made per task, turning what was a single inference into dozens. Reasoning models — GPT-5-class, Claude Opus, Gemini Pro — consume three to ten times more compute per query than their predecessors because they think longer before responding. And consumer AI usage continues to grow, competing for the same hyperscaler capacity that enterprises need.

Gartner, IDC, IEA, and the Uptime Institute all project continued demand growth through 2030. The debate is about pace, not direction.

Supply side

Supply is constrained across three independent bottlenecks, as described in Chapter 2: silicon capacity, data center construction, and power generation. Each has its own time constant, and they do not respond to price the way typical commodities do. You cannot pay extra to have a transmission line built in six months. You cannot spin up an advanced-node fab in two years. You cannot commission a gigawatt of new baseload generation quickly, regardless of budget.

There is a counterargument that hyperscalers are over-building and that actual demand will plateau. ITIF, for example, has argued that only a third of announced data center capacity will actually be constructed and that hyperscaler capex may fall by half in 2026. We find this argument partially persuasive: yes, some announced projects will slip or cancel. But it does not change the constraint-shaped nature of supply. Even a reduced buildout pace leaves supply tighter than demand, because the demand side is accelerating, not plateauing.

Where the curves intersect

The forward-looking question is what happens where the demand curve crosses the supply curve. We do not yet have direct public evidence of widespread inference capacity rationing at hyperscalers — the current signals are indirect: rate limits that appear during peak periods, enterprise customers reporting variable pricing on multi-year commits, and public disclosures from providers about prioritizing their largest accounts

for scarce GPU capacity. These are precursors of rationing, not proof of it. But the structural setup — supply-constrained by physical bottlenecks, demand still accelerating — is the classic signature of a market where price and availability will both tighten. Treating rationing as a plausible risk scenario, rather than an inevitability, is the appropriate posture for a security leader planning through 2029.

The more confident claim is about pricing. Frontier model pricing will hold or drift upward; mid-tier model pricing will continue to decline, but more slowly than the 2023-2025 trajectory suggested. The easy cost-reduction paths — larger batches, cache discounts, open-weight substitution for simple tasks — will be exploited; what remains is expensive. This is visible in present data, not extrapolated from it.

Organizations that have already optimized their AI usage will be fine in either scenario. Organizations still architected around the "feed everything to the LLM" pattern are exposed to both: progressively painful pricing if the market merely tightens, and operational disruption if rationing does materialize.

~$7T

McKinsey's estimate of capital investment required to meet projected data center energy demand by 2030

Timeline

We expect the following rough timeline based on current industry indicators:

- **2026** Frontier token prices stabilize. Enterprise customers begin reporting capacity rationing at peak. First serious SOC AI cost cuts begin at large enterprises. In October, five national cyber security agencies publish joint guidance on managing AI token consumption for operational resilience.
- **2027** Gartner's 40% data center restriction forecast either materializes or is materially revised. CFOs begin auditing AI spend line items in earnest. Organizations with per-token cost exposure start renegotiating vendor contracts regardless of whether rationing appears, because pricing pressure alone is enough to force the issue.
- **2028** New data center capacity and modest grid expansion begin to relieve the worst regional bottlenecks. Prices still elevated. Empathetic-processing architectures are visibly cheaper than AI-first alternatives.
- **2029 and beyond** Long-lead power and fab investments start to pay off. Capacity loosens. But by then, the efficient architectures are entrenched and the inefficient ones have been re-architected or abandoned.

19 · Compounding advantage

The compounding advantage of empathetic processing has three dimensions: cost avoidance, capacity resilience, and architectural flexibility. Each one becomes more valuable over time.

Cost avoidance compounds

The worked example in Chapter 16 showed a single year of operation: $2.58 million of annual savings, a 79% cost reduction. Project that across five years — with modest data growth, AI token prices holding roughly steady, and the customer's existing SIEM contract lapsing at normal renewal intervals — and the cumulative

saving is meaningful but sensitive to assumptions. A reasonable mid-case estimate lands in the eight-figure range for a single mid-sized enterprise over that horizon; scenarios with faster data growth, more aggressive SIEM renegotiation, or heavier AI adoption push the figure higher. At the scale of a large enterprise or a multi-tenant MSSP, the savings compound faster but the uncertainty compounds with them.

The specific number matters less than the direction. What the model reliably shows — and what survives sensitivity analysis across plausible parameter ranges — is that costs in the raw-data architecture grow with data volume, while costs in the empathetic-processing architecture do not. That structural difference, not any single savings figure, is the argument this paper is making.

This is not one-time savings. It is recurring, growing, and compounding with data volume growth.

Capacity resilience compounds

Less obvious but equally important: organizations with empathetic processing in place are less exposed to inference capacity pressure, whether it manifests as rate limits during peak demand, as preferential allocation to providers' largest accounts, or as outright rationing should supply constraints intensify. A lean AI workload is easier to service than a heavy one. An AI SOC spending $10,000 per month on inference has far more provider flexibility, pricing leverage, and fallback options than one spending $100,000 — because the former can be served by mid-tier models or even open-weight alternatives, while the latter requires frontier capacity at premium pricing.

Empathetic processing is not just a cost optimization. It is an operational hedge against whatever form of supply pressure does emerge.

Architectural flexibility compounds

A security program built around empathetic processing can change AI vendors freely. The artifact stream leaving Conductor is structured JSON (or the original log line), delivered to the SIEMs, data lakes and object stores Conductor connects to, including Splunk, Microsoft Sentinel, CrowdStrike and OpenSearch, or over syslog. From there it can feed the AI SOC platform or frontier LLM you choose. If OpenAI rate-limits your account, you switch to Anthropic. If Anthropic prices go up, you test Gemini. If Gemini's quality drops, you fail back. The investment in curation is portable; the investment in a specific AI vendor's ecosystem is not.

Organizations that bet heavily on a single AI SOC platform, or a specific LLM vendor's ecosystem, have much less optionality. They will find renegotiating harder, switching harder, and walking away impossible. The WitFoo layer, by design, keeps options open.

Security operations at a New Zealand insurer.

"We have deliberately brought security into our AI work from the outset, so it develops alongside the technology rather than responding afterwards. AI accelerates existing challenges across data protection, insider risk, identity, hardening and secure software development, but it also gives us the opportunity to move beyond reacting to individual alerts and analyse larger datasets for connected risks and attack patterns. We are also exploring greater flexibility across AI models, selecting the right model for the task rather than concentrating capability, cost and dependency in one place. That makes optimising the data submitted, using deterministic automation where reasoning is not required, and reserving advanced AI capacity for high-value security analysis essential parts of the design."

Rob Kehl, Senior Security Operations Specialist, Tower Insurance

The three-way compounding

Cost avoidance compounds annually with data growth.

Capacity resilience compounds with every supply tightening.

Architectural flexibility compounds with every AI vendor shift.

All three point the same direction: the advantage of empathetic processing grows, not shrinks, as the AI market matures.

20 · Recommendations for 2026 through 2029

The economic analysis produces a clear set of operational recommendations for security leaders planning capacity and architecture through the end of the decade. These apply whether the organization is currently running a traditional SIEM, an AI-first platform, a data lake, or a hybrid architecture.

For CISOs

Audit your AI cost exposure. Every contract with a per-token, per-alert, per-event, or per-GB pricing mechanism is a variable-cost time bomb if data growth exceeds your budget growth. Categorize every line item by its cost-scaling behavior. Model the three-year trajectory under realistic data growth assumptions. Any line item that scales faster than your budget is a candidate for replacement or mitigation.

Insist on architectural separability. Any vendor asking you to send raw data directly into their analysis layer is asking you to surrender the single most valuable optimization lever in the stack. Keep the data collection, curation, and analysis layers separable. An empathetic processing layer in front of your analysis stack gives you room to maneuver — change analysis vendors, change AI providers, adapt to capacity shocks — without re-architecting your entire data plane.

Put the SOC in the top tier. The agencies' joint guidance recommends separating token allocations by business impact, with cyber defense, incident response, and threat intelligence in the highest tier. Classify your AI workloads the same way, test that the Tier 1 workloads hold up under the demand you expect before you add lower-priority uses, and keep the non-AI capabilities you would need if AI services were degraded. Set alerts

as consumption approaches its limits (the guidance's example is an information alert at 70 percent of quota, a warning at 80, and a critical alert at 90), and feed AI token telemetry into your SIEM, correlated with identity, cloud, API, and system logs. The guidance notes that an unusual increase in token consumption can be an indicator of data exfiltration; investigate it as you would any other anomaly.

For CFOs

Stress-test every security vendor's pricing model. Ask for a three-year price curve assuming data volume doubles. Many vendors will refuse or quote wildly divergent numbers — that is itself information. Per-appliance, per-unit, per-seat, and fixed-capacity models are the predictable ones. Volume-scaled, event-scaled, and token-scaled models are the ones that surprise you on renewal.

Treat AI inference as a cost center requiring governance, not as a feature that automatically delivers ROI. Every dollar of LLM spend should tie to a specific analytic outcome that a deterministic system could not produce. The default assumption should be that deterministic logic handles the common case, with AI reserved for the hard cases. This is how every other expensive cognitive resource in an organization is managed.

For channel partners and solution architects

Lead with the cost conversation. The CISO and CFO are both newly receptive to architectures that insulate them from AI price exposure. Empathetic processing — and the WitFoo pricing model underneath it — is a direct response to a problem they are actively trying to solve. The old pitch ("make your SIEM better") still works; the new pitch ("shield your AI budget") has become even more compelling.

Structure deployments for maximum protection against volume growth. Where possible, deploy Conductor Pro rather than Lite — the incremental cost is small relative to the protection against future data growth. Where possible, deploy Analytics at HQ and Conductors at sites, rather than forwarding everything to a central cloud SIEM. Where possible, retain data in the customer's environment rather than in a vendor's cloud, to avoid surprise egress or retention charges.

For the industry

The AI-in-security conversation of 2026 to 2029 will be shaped by the organizations that recognize the cost collision early. A small number of enterprises — and the security vendors, MSSPs, and integrators who advise them — will be architected correctly for the tightening AI market. The rest will learn the hard way. Our recommendation is to be in the first group.

> **The asymmetric bet**
>
> Empathetic processing costs nothing if AI compute becomes plentiful and cheap — you save on SIEM, storage, and analyst time regardless.
>
> Empathetic processing wins dramatically if AI compute stays constrained and expensive — which every forecast we trust says is the more likely outcome.
>
> This is an asymmetric bet with a large upside, a small downside, and high confidence in the direction of travel. In risk terms, it is close to free.

The next thirty days

Deploy, measure, decide. Put a Conductor in front of one high-volume source during the 15-day free trial, and measure what reaches your SIEM, your data lake, and your AI tools before and after. The result is your own number, from your own data.

CONCLUSION

The next four years will separate the architectures.

The cybersecurity industry has been telling itself a story about AI: that models will keep getting cheaper, that compute will keep getting more plentiful, that the architectural patterns of 2024 and 2025 will smoothly mature into the autonomous SOC of 2027 and beyond. Every piece of evidence available in early 2026 suggests the story is incomplete. Frontier prices are stabilizing. Capacity pressure is visible in rate limits, preferential allocation, and the early signals of rationing. Power is the binding constraint on supply, and power does not care about product roadmaps. By October 2026 the national cyber security agencies of five countries were advising organizations to manage AI token consumption as an operational-resilience risk, and to put cyber defense at the front of the queue.

Security is among the hardest places this collision will land, because security generates the largest data volumes, demands the most per-event analysis, benefits most from agentic workflows, and has the strictest retention requirements. The “feed everything to the LLM” architecture that dominates today’s product marketing is an economic dead end for enterprises that will need to scale into an AI-constrained world.

Empathetic processing — the discipline of using security-aware, deterministic logic to curate, deduplicate, normalize, and enrich data before any expensive model sees it — is the only architectural answer that closes the gap without giving up AI’s genuine contributions. It is not an alternative to AI. It is what makes AI economical at enterprise scale.

WitFoo has been building this layer for seven product generations. Our ProtoGraph deduplication, adaptive parsing, and unified artifact output give downstream AI systems — whether our own Analytics, a third-party AI SOC platform, or a customer-built frontier LLM workflow — dramatically smaller, dramatically richer, dramatically cheaper inputs. Our per-appliance pricing model means the cost of that curation layer does not scale with data volume. The two together shield customers from the worst of what is coming in the AI compute market. The innovations described in this paper are protected by seven provisional patent applications filed with the USPTO in July 2026 (patent pending); a provisional application establishes a priority date, is not an issued patent, and no assurance can be given that any patent will be granted or will provide the scope sought. Together they make WitFoo the deterministic pane of glass for cybersecurity: every record processed, every calculation defensible, where the alternatives offer probabilistic guesses.

The next four years will separate security programs that treat AI as a precision tool from those that treat it as a firehose. Only the former will be affordable. Only the former will have the capacity they need. Only the former will keep the optionality to change direction as the market shifts.

This is not a product pitch. It is an economic observation, substantiated by public supply-side data and a pricing model that any CFO can verify. The conclusion follows the evidence: empathetic processing, as a category and as WitFoo implements it, is the economic imperative of the next decade in cybersecurity.

WitFoo — Cybersecurity for Collective Defense.

How this paper was produced

WitFoo uses artificial intelligence tools in producing its published research, and we think readers are entitled to know how. This section explains what the tools did, what they did not do, and where responsibility sits.

What AI tools were used for. We used AI tools in three ways. First, research: gathering candidate sources, summarising long documents, and surfacing data we then went and read for ourselves. Second, fact-checking: cross-referencing figures, dates, names, and quotations against independent sources and flagging anything that could not be corroborated. Third, editing: tightening prose, checking consistency of terminology and units, and catching typographical and formatting errors.

What AI tools were not used for. The argument in this paper is ours. The thesis, the structure, the selection of evidence, the interpretation of that evidence, and every conclusion drawn from it were decided by WitFoo's authors. No statistic, quotation, or citation appears in this paper because a model produced it; each was traced to a primary or reputable secondary source and verified by a person before publication. Where we could not verify a claim to our satisfaction, we removed it or labelled it as unverified in the text.

Where responsibility sits. WitFoo is the author of this paper and is solely responsible for its accuracy, its reasoning, and its conclusions. Our use of AI tools does not transfer, dilute, or share that responsibility. If an error appears in this paper, it is our error. Corrections should be directed to press@witfoo.com, and we will publish an erratum where one is warranted.

Whose views these are. The positions taken in this paper are WitFoo's, informed by our own product development, customer deployments, and research programme. They are not the views of any AI vendor, model provider, or tool. The text, analysis, figures, and any original frameworks or terminology presented here are WitFoo's intellectual property and are protected accordingly.

Why we disclose this. WitFoo builds software that applies analytics and machine learning to security operations data. We hold that AI is a useful instrument when a human sets the question, checks the answer, and owns the result, and a poor one when it does not. Publishing under that standard, and saying so, is how we hold ourselves to it.

ABOUT WITFOO

About WitFoo

WitFoo builds security operations software that reads all of an organisation's security data, reduces telemetry volumes by more than 90% without losing evidence, and shares what it finds with the people who must act on it, from analysts and boards to law enforcement. Founded in 2016, WitFoo, Inc. is headquartered in Dunwoody, Georgia, and sells its Conductor, Analytics and Reporter products through channel partners. Its New Zealand subsidiary, WitFoo Limited, is based at 47 Salisbury Street in Christchurch Central, where the company is building its international commercial operations. Cybersecurity for Collective Defense.

Press inquiries: press@witfoo.com. Investor and shareholder inquiries: charles@witfoo.com.

APPENDIX A

WitFoo pricing reference

All WitFoo 7th Generation pricing is per appliance, per annum, with unlimited data rates within each appliance's capacity envelope. No volume charges, no per-alert fees, no token pass-throughs. All 7G products are Generally Available: Analytics and Reporter reached General Availability on 30 June 2026; Conductor was already GA. Pricing current as of October 2026; subject to change.

Standard list pricing

Product	Lite	Pro	Max
Conductor — USD	$15,000	$40,000	$60,000
Conductor — NZD	$19,500	$52,000	$78,000
Reporter — USD	$20,000	$50,000	—
Reporter — NZD	$26,000	$65,000	—
Analytics — USD	$25,000	$75,000	$120,000
Analytics — NZD	$32,500	$97,500	$156,000

Government, non-profit, and education pricing (15% discount)

Product	Lite USD	Pro USD	Max USD
Conductor	$12,750	$34,000	$51,000
Reporter	$17,000	$42,500	—
Analytics	$21,250	$63,750	$102,000

Volume discounts

Volume discounts have applied since 24 July 2026. Quantities are counted per product line within an order, and editions count together: Conductor Lite, Pro, and Max all count toward the Conductor quantity. The first appliance of each product line is always at full published price; every additional appliance of that line takes the discount set by the line's total quantity in the order. Where an order mixes editions of one product line, the highest edition present is the appliance at list. For government, non-profit, and education customers, the 15% discount applies first, then the volume tier.

Order quantity (per product line)	First appliance	Each additional appliance
1	List	—
2–5	List	30% off list
6–10	List	50% off list
11+	List	65% off list

Key licensing rules

- Reporter and Analytics bundle a Conductor at no extra cost (Lite bundles Conductor Lite; Pro/Max bundle Conductor Pro).
- Each appliance instance requires its own license, including HA pairs and Analytics data nodes.
- CyberGrid data synchronization is available at all Analytics tiers.
- Console is free and recommended for deployments with 3 or more appliances.

Hardware requirements

Role	CPU (min)	RAM (min)	Disk (min)	Recommended
Conductor	4 cores	8 GB	220 GB	8 CPU / 16 GB / 500 GB
Console	4 cores	8 GB	220 GB	4 CPU / 8 GB / 220 GB
Analytics / Reporter	8 cores	12 GB	220 GB	16 CPU / 32 GB / 1 TB

Supported platforms: VMware, Hyper-V, QEMU/KVM, AWS, Azure, Google Cloud, and bare metal running Ubuntu 24 or RHEL 10. See github.com/witfoo/witfoo-appliances for current benchmarks and deployment guides.

APPENDIX B

References

The macroeconomic and technical claims in this paper draw on a range of publicly available industry research, vendor pricing disclosures, and WitFoo operational data. Key sources are grouped by topic below. Market prices and forecasts are as of April 2026 unless dated otherwise; WitFoo pricing is as of October 2026. All are subject to change.

Government guidance

- Canadian Centre for Cyber Security, with the Cybersecurity and Infrastructure Security Agency (US), the Australian Signals Directorate's Australian Cyber Security Centre, and the National Cyber Security Centres of New Zealand and the United Kingdom — *Managing AI token consumption for operational resilience* (ITSM.00.125), October 2026. TLP:CLEAR. https://www.cyber.gc.ca/en/guidance/managing-ai-token-consumption-operational-resilience-itsm00125 ; also published at https://www.ncsc.govt.nz/protect-your-organisation/managing-ai-token-consumption-for-operational-resilience/ , https://www.cyber.gov.au/publication/managing-ai-token-consumption-for-operational-resilience and https://www.cisa.gov/resources-tools/resources/managing-ai-token-consumption-operational-resilience

Data center capacity and energy constraints

- Uptime Institute — 2026 Predictions Report; Annual Outage Analysis; power constraint forecasts. https://intelligence.uptimeinstitute.com/resource/five-data-center-predictions-2026 ; https://intelligence.uptimeinstitute.com/resource/annual-outage-analysis-2026 ; https://intelligence.uptimeinstitute.com/resource/uptime-institute-global-data-center-survey-2026
- International Energy Agency — Energy and AI Analysis; data center electricity demand projections through 2035. https://www.iea.org/reports/energy-and-ai ; https://www.iea.org/reports/key-questions-on-energy-and-ai
- Gartner — Hype Cycle for Emerging Technologies; AI data center restriction forecasts. https://www.gartner.com/en/documents/8163629 ; https://www.gartner.com/en/newsroom/press-releases/2024-11-12-gartner-predicts-power-shortages-will-restrict-40-percent-of-ai-data-centers-by-20270
- Congressional Research Service — January 2026 hyperscale data center report. https://www.congress.gov/crs-product/R48646
- Belfer Center for Science and International Affairs — AI, Data Centers, and the U.S. Electric Grid (February 2026). https://belfercenter.org/research-analysis/ai-data-centers-us-electric-grid
- CBRE — North America Data Center Trends, H1 2025 and updates. https://www.cbre.com/insights/reports/north-america-data-center-trends-h1-2025 ; https://www.cbre.com/insights/books/north-america-data-center-trends-h2-2025 ; https://www.cbre.com/insights/books/north-america-data-center-trends-h1-2026
- McKinsey & Company — Global data center capital investment projections. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-cost-of-compute-a-7-trillion-dollar-race-to-scale-data-centers

- ITIF — Four Reasons New AI Data Centers Won't Overwhelm the Electricity Grid (April 2026, counter-perspective). https://itif.org/publications/2026/04/07/four-reasons-new-ai-data-centers-wont-overwhelm-the-electricity-grid/

LLM pricing and token economics

- Silicon Data — 2026 LLM Cost Per Token Practical Guide. https://www.silicondata.com/blog/llm-cost-per-token
- AI Cost Board — LLM API Pricing Comparison 2026. https://aicostboard.com/blog/posts/llm-api-pricing-comparison-2026
- Price Per Token — LLM Pricing Trends 2026. https://www.pricepertoken.com/pricing-history
- CostGoat — LLM API Pricing Comparison (April 2026). https://costgoat.com/compare/llm-api
- Prompt Engineer Collective — LLM API Pricing 2026 (20+ Models Compared)

AI in security operations

- Dropzone AI — Agentic SOC platform; published pricing and investigation metrics. https://www.dropzone.ai/pricing ; https://www.dropzone.ai/ai-soc-analyst
- Intezer — Top 15 AI SOC Tools for 2026. https://intezer.com/guides/ai-soc
- D3 Security — Best AI SOC Platforms 2026 comparison guide. https://d3security.com/blog/ai-soc-platforms-2026/
- Elastic — AI for Security Operations and SOC Teams. https://www.elastic.co/security/ai
- MDPI — AI-Augmented SOC: A Survey of LLMs and Agents for Security Automation. https://doi.org/10.3390/jcp5040095
- Gartner — AI SOC category analysis and maturity projections. https://www.gartner.com/en/documents/7962373 ; https://www.gartner.com/en/documents/7075998

WitFoo product documentation

- WitFoo 7th Generation Product & Pricing Reference Guide (internal)
- WitFoo 7th Generation Selling Guide (internal)
- WitFoo 7th Generation Deployment Quoting Guide (internal)
- docs.witfoo.com — public product documentation
- https://www.witfoo.com/documents/Empathetic-Processing-Whitepaper.pdf — Empathetic Processing in Cybersecurity: A Human-Centric Data Analytics Paradigm (Charles D. Herring, November 2025)
- https://www.witfoo.com/research/precinct-6-dataset/ — the Precinct 6 Cybersecurity Dataset: more than 114 million sanitized security events captured live from five organizations in 2024, released under Apache 2.0 and produced in collaboration with the University of Canterbury (UC) | Te Whare Wānanga o Waitaha Computer Science and Software Engineering department. Also on Hugging Face: https://huggingface.co/datasets/witfoo/precinct6-cybersecurity-100m
- github.com/witfoo/witfoo-appliances — hardware benchmarks and deployment guides

Quoted in this paper

- Tom Cross, cybersecurity expert and WitFoo advisor. https://www.linkedin.com/in/tom-cross-/
- TK Keanini, cybersecurity technologist and author of tkonsecurity.com. https://tkonsecurity.com/
- Luke Taylor, Chief Executive Officer, Scientific Software and Systems. https://www.linkedin.com/in/luke-taylor11/
- Gopal Padinjaruveetil, KnowNothingCISO. https://www.linkedin.com/in/gpadinjaruveetil/
- Rob Kehl, Senior Security Operations Specialist, Tower Insurance.
- Leilani Lauger, Higher Education CISO.
- Josh O'Sullivan, Chief Technology Officer, Ardalyst. https://www.linkedin.com/in/joshosullivan/
- Thomas Whang, Chief Technology Officer, Impelix.

WitFoo

Cybersecurity for Collective Defense.

witfoo.com · Press: press@witfoo.com · Investors and shareholders: charles@witfoo.com

WitFoo uses artificial intelligence tools to assist with research, fact-checking, and editing in the preparation of this paper. All analysis, conclusions, and opinions are WitFoo's own. WitFoo reviewed every claim, selected every source, and takes full responsibility for the content. Nothing in this paper was generated without human direction and review. The views expressed represent WitFoo's perspective.

© 2026 WitFoo, Inc. This paper is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/. You may share and adapt it for any purpose, including commercially, if you credit Charles D. Herring, The AI cost collision in cybersecurity (WitFoo, Inc., 2026), link to the license and say whether you made changes.

The perspectives quoted in this paper are on-record and approved for distribution by each person quoted. The government guidance cited is public; the agencies that published it did not review this paper and do not endorse WitFoo or any product.

The AI cost collision in cybersecurity. Charles D. Herring, WitFoo, Inc. Version 3.9, October 2026 (first drafted April 2026). Dollar figures are US dollars unless marked NZD.