Aegis AI - EPP/EDR/XDR Industry Disruptor

Where comparable endpoint protection, detection and response products cannot compete:

- **AI on the protected machine.** Two Llama language models on the endpoint's own CPU explain every incident and, in the finished product, decide on it. Competitors run their AI in their own cloud, for an analyst.
- **You take part in training the AI.** As and when you need to, you add your own knowledge -- your policies, your environment, incidents you have reviewed -- to both models on the easy LoRA training page. It only adds to what they detect: it builds on the extensive fine-tuning AegisAI provides and keeps current with regular updates.
- **You take part in defining threats.** Your own Sigma, YARA and signature rules sit beside AegisAI's and define the threats and incidents that matter on your machine or in your company, breaking out of the industry's one-size-fits-all standard.
- **Explained to you, air-gapped by design.** A plain-English storyline mapped to MITRE ATT&CK for every incident; full capability with no network, and nothing sent to a vendor.
- **Yours to shape, open to audit.** Choose the engines, their order and how far the response may go on its own. Every verdict traces to an open rule; every action is signed and reversible.

A perpetual licence, with no compulsory retainer. An optional retainer, from as little as **USD 5 per month per machine**, delivers new rules, definitions, signatures and AI model fine-tuning files. They are added to your own, never overwrite them, and always give priority to what YOU consider important and actually risky for your system's use and environment -- or for your SME's data, policies, security and privacy. The leading AI-driven platforms charge about USD 180-230 per endpoint per year.

AEGISAI LOCAL - ENDPOINT PROTECTION THAT NEVER LEAVES THE MACHINE

Detect, respond and recover -- entirely on the device, explained in plain English.

AegisAI is a fully local endpoint-protection platform for Windows. It watches the machine through signed, tamper-evident telemetry, catches threats with layered deterministic detectors and on-device machine learning, and explains every behavioural incident with two Llama language models running on the CPU. Nothing is uploaded, nothing depends on a cloud, and it works the same on an air-gapped workstation as on a connected one.

100% on-device | offline / air-gap ready | no data retention | two local Llama models | you take part in training the AI

What AegisAI does today

This build is a working prototype of AegisAI's core. It already runs the complete chain from kernel telemetry to an explained, contained incident on an ordinary PC -- and it is deliberately strict about one thing: in this prototype the rules decide, and the language models only explain.

IN THE FINISHED PRODUCT

In the final product the fine-tuned dual-LLM pipeline will also make decisions, based on its own explanations and analyses of what a file or a process does or may do. It will discover 0-day threats that nothing else available today can detect, and it will detect threats defined by you: by the particulars of your computer's set-up and your own policies, or, in AegisAI Local - SME Edition, by your company's security policies and practices.

That is the step from a product that tells you what happened to one that understands what is happening on your machine, judged against your own rules, and acts on it -- with the same safety floor, audit trail and reversibility that this build already enforces on every action.

A multi-pronged detection core

Curated SigmaHQ behavioural rules over ETW and Sysmon telemetry, YARA rules on files and in process memory, ClamAV signatures, and two ONNX scorers -- a supervised behaviour model and an unsupervised novelty model. A threat missed by one layer can be caught by another, and agreement across layers raises an incident's severity. In the final product you choose which engines examine the files during a scan, the dual-LLM reasoning pipeline among them, and the order in which they are used.

Two Llama models, on the CPU

A stock Llama 3.2 1B gives every behavioural incident a first read within seconds; a deterministic router ranks the work; a Llama 3.2 3B fine-tuned for AegisAI writes the authoritative explanation, mapped to MITRE ATT&CK. Analyst-grade narratives, produced without a byte leaving the machine.

Response with a safety floor

Malicious files are quarantined into a protected, restorable store; malicious processes are suspended, not killed, so their memory survives for analysis. A safety floor refuses to touch critical system processes or validly signed binaries, every action is audited, and every one can be undone.

Incident controls, checked before they are offered

Each incident offers only the actions that can work on this machine at that moment, decided by the engine and checked again every few seconds. A button for a process that has exited, a signed file or a protected system folder is never shown; one that applies but not yet is greyed out, with the reason in its tip. A running process can be paused, resumed, asked to close, force-killed, or force-killed with its whole process tree -- children first, and the whole tree refused if any member is one Windows needs. A file can be quarantined and restored to its original path, or deleted together with the files the program wrote, and every file involved is listed with its full path. A file can also be declared safe from its incident, even one no engine can read: what was done to it is reversed -- the quarantined copy restored, the paused program resumed -- and its incidents are removed from the journal, the action audit and the logs as if they never happened. A signed program that a behavioural rule matched cannot be declared safe, so a trusted system tool can never be cleared by accident.

Sharper and faster detection

157 curated rules, including AegisAI's own rule for an unsigned program started from a user's AppData Roaming folder, a common staging place for malware. Process-tree corroboration is bounded to the incident's own time window, a vendor-signed program in its own Program Files folder is no longer mistaken for a renamed system tool, and a cache of signature verdicts makes a repeat quick scan about three times faster. A system scan names the engine and the rule behind every finding and every file it could not examine. A file rescanned successfully or declared safe leaves those lists, and a locked system file declared safe by its path is passed by while it stays unreadable, then examined like any other once it can be read.

Honest status, built-in attribution

The product reports reduced capability rather than painting a green tick over a degraded system, and every detection rule's author is credited wherever its matches appear, as the Detection Rule License requires.

Prototype requirement: Microsoft Defender and Smart App Control must be completely off

Windows hands the role of endpoint-protection product from Microsoft Defender to another product only when that product registers with Windows Security Center, and registration requires a Microsoft-approved driver: an Early Launch Anti-Malware driver signed by Microsoft through its Microsoft Virus Initiative. This prototype has no such Microsoft driver, so Windows does not recognise AegisAI as the machine's protection and will not stand Defender down in its favour.

That generates the need for Microsoft Defender to be completely disabled for the prototype to be demonstrable. With Defender running, the two would race for every file, and Defender would remove threats before AegisAI saw them -- a demonstration would show Defender's response, not AegisAI's. AegisAI changes nothing about Defender itself: the installer warns if Defender is running, and the readiness check (aegis-service assess) reports it as a blocker.

Smart App Control (SmartScreen on Windows 10) must be completely off as well, for the least disruption and the easiest installation: this prototype is not code-signed, and Smart App Control stops unsigned programs from starting. The installer checks it and does not install while it is on.

Step-by-step PDF guides for turning both off, and PowerShell scripts that make the process almost completely automatic, are downloadable together with the installer.

Anyone who wants to test this advanced prototype should do it in a sandboxed environment, such as a virtual machine, that holds no sensitive or personal information. The scripts are provided as helpers, so that the prototype can be tried with minimum risk.

Never try to install this prototype on your regular computer or disable Windows Defender and/or SmartAppControl (SmartScreen on Windows 10) on a real non-sandboxed computer/environment.

ALREADY BUILT - SHIPS IN THE FINAL PRODUCT

AegisAI's own kernel network-enforcement driver

AegisAI has its own Windows Filtering Platform (WFP) callout driver, developed and working. It sits in the Windows network stack and makes a block-or-permit decision on every connection, so a malicious process can be cut off before its first packet leaves the machine -- enforcement, not just observation.

It is left out of this prototype for one reason only. Windows loads a kernel driver only when Microsoft has signed it, and Microsoft's attestation signing requires an EV code-signing certificate, which is not yet in place. Running the driver unsigned would mean switching the machine into test-signing mode -- Secure Boot off, and a permanent "Test Mode" watermark on the desktop -- and that is no way to present a security product. Until the driver is signed, this prototype observes network connections through Sysmon; the final product ships the driver and enforces.

developed and working | blocks before a packet leaves | awaiting a driver-signing certificate

You take part in training the AI: model fine-tuning on your own machine

Teach the two local language models your own environment and your own security policy, on this machine. AegisAI's Model fine-tuning page lets their owner -- the person at the machine or, in AegisAI Local - SME Edition, the company's own security team -- take part in training both language models, as and when it is needed. What the owner adds builds on the extensive fine-tuning AegisAI provides and keeps current with regular updates: it adds to what the models can detect and never replaces it. No endpoint-protection product on the market today offers its customers anything like it.

WHAT IT IS FOR

Out of the box both models know Windows threats in general and nothing about your organisation: which administration tools you sanction, which scripts IT runs at night, where your own software lives, which practices your policy forbids and whom to call. So an explanation can only say what happened, not whether it fits the way you work.

Fine-tuning teaches them that knowledge. An explanation can then say "this is the nightly backup job your policy allows in section 4.2, started from its approved folder" instead of "an unknown script ran with administrator rights", and the first read of an incident reflects what is routine here. You spend less time on behaviour you have already sanctioned, and every explanation is written against your own rules.

LoRA (low-rank adaptation) does this without retraining a model. It learns a small adapter file, tens of megabytes, that sits on top of the unchanged base model. An adapter can be tested beside the one in use before it is switched on, and switched off again by going back to the previous file.

In the finished product the tuned models do more than explain: the pipeline also decides, and it detects threats defined by your own set-up and policies -- or by your company's.

Tier 1: triage

the fast first read of every incident

An adapter teaches it your vocabulary and what is routine on your machines, so its first read, and the order in which the router hands incidents to the deep model, reflect your environment.

Tier 2: deep explanation

the authoritative explanation

Your knowledge is added to the AegisAI training, not put in its place: the new adapter learns from the AegisAI examples and yours together, so the security knowledge the model already has is kept.

How it works

1

Bring your knowledge

Upload a personal or company security policy, or pick incidents you have already reviewed and write the explanation each one should have had.

2

Turn it into training examples

AegisAI parses each item into lines of a JSONL training file, one example per line, in the exact form each model reads when it works: the same fixed instructions, an event window drawn by the product's own renderer, and the answer the model should give. Training and use see the same bytes, which is what makes an adapter work.

3

Review and choose

Every extracted item is listed with the place it came from. You edit its wording, keep it or drop it, and choose which model receives it: the triage model, the deep model or both.

4

Train

A LoRA adapter is trained for each chosen model on its full-precision base, on this PC or on a training machine you name. Only the adapter's weights are learned; the base model does not change.

5

Check before use

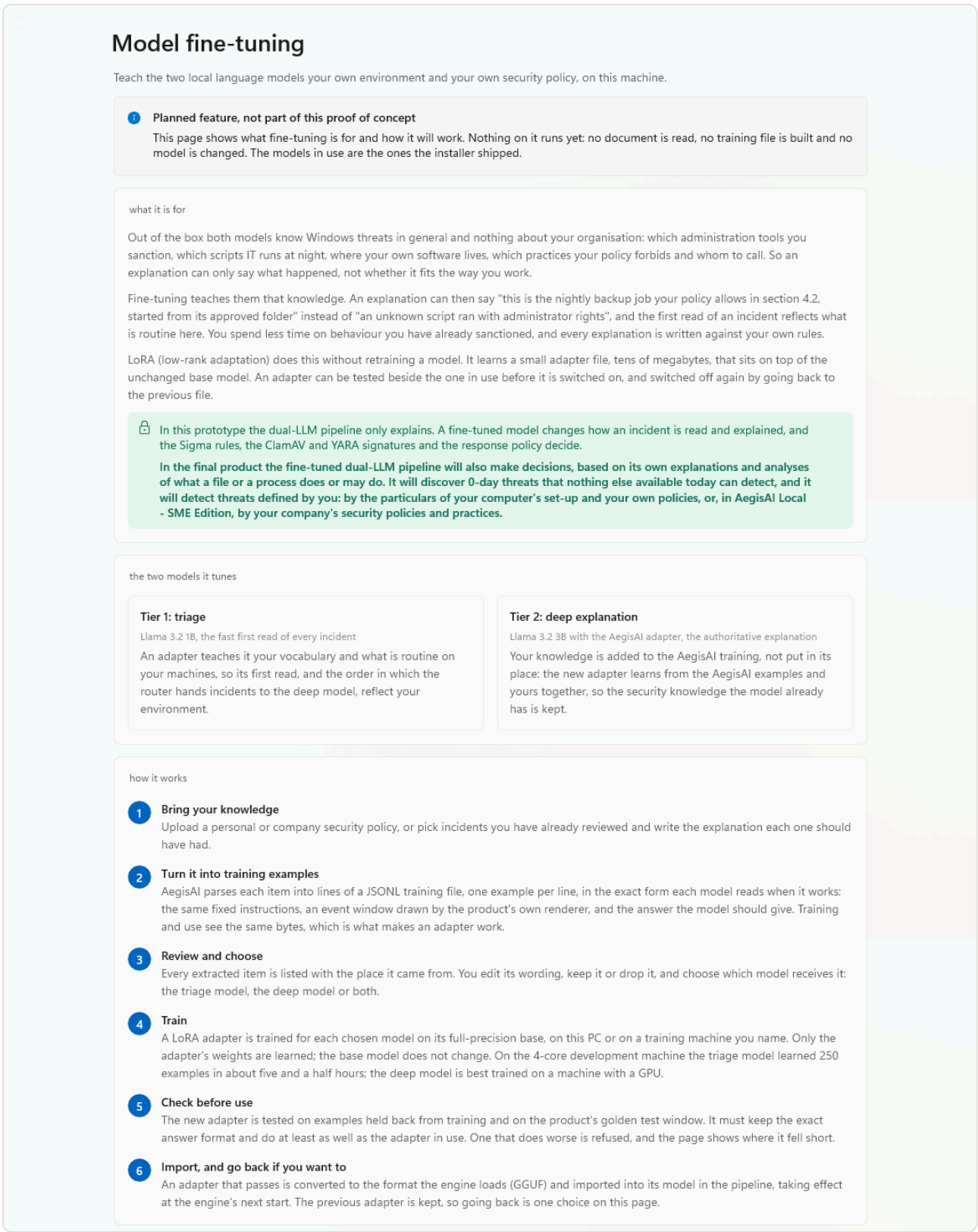
The new adapter is tested on examples held back from training and on the product's golden test window. It must keep the exact answer format and do at least as well as the adapter in use. One that does worse is refused, and the page shows where it fell short.

6

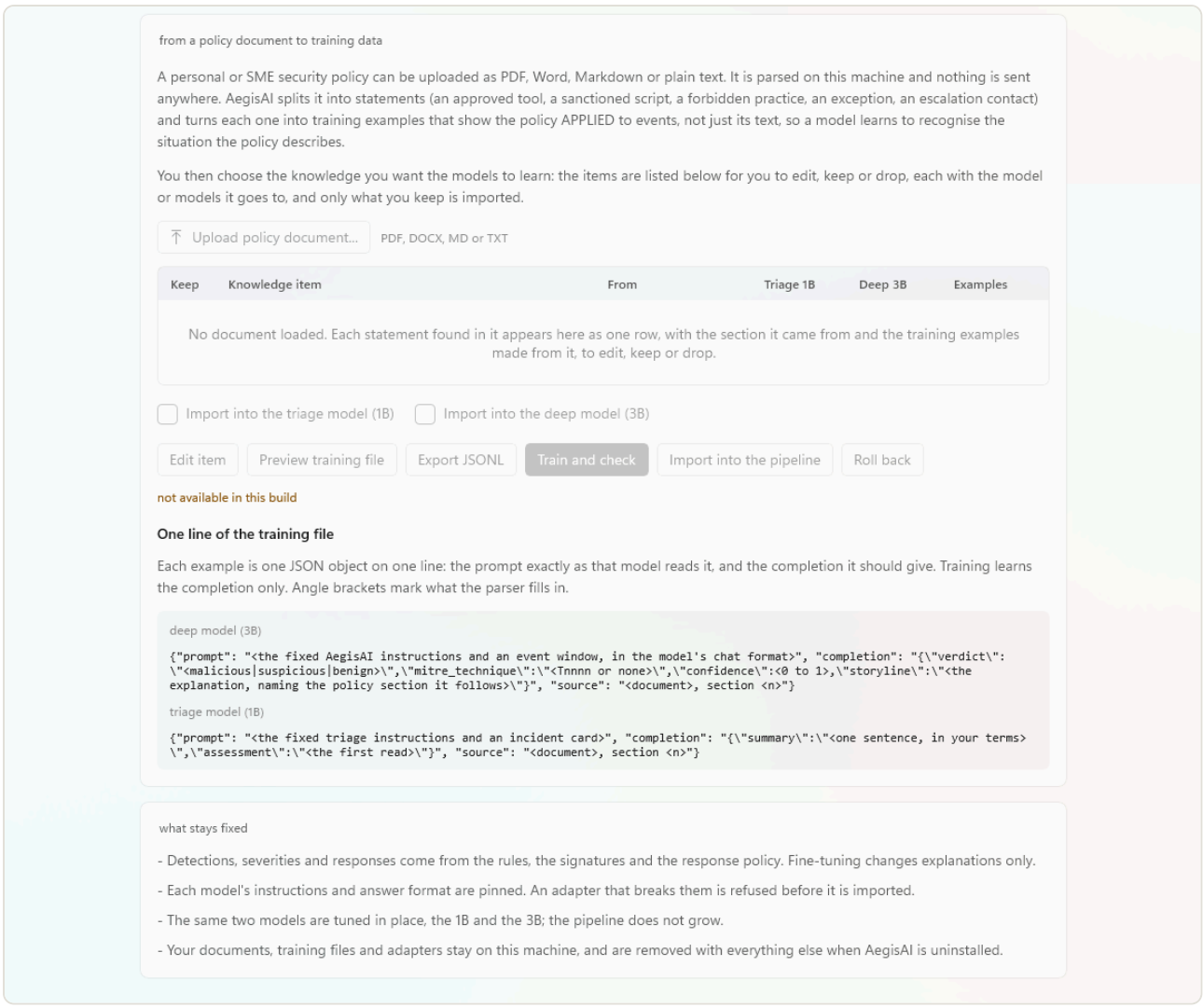
Import, and go back if you want to

An adapter that passes is converted to the format the engine loads (GGUF) and imported into its model in the pipeline, taking effect at the engine's next start. The previous adapter is kept, so going back is one choice on this page.

The Model fine-tuning page -- from a policy document to a tested, reversible adapter, without anything leaving the machine.



The Model fine-tuning page in the AegisAI Management app, scrolled top to bottom (1 of 2): what it is for, the two models it tunes, and how it works.



The Model fine-tuning page (2 of 2): from a policy document to training data, one line of the training file, and what stays fixed.

From a policy document to training data

A personal or SME security policy can be uploaded as PDF, Word, Markdown or plain text. It is parsed on this machine and nothing is sent anywhere. AegisAI splits it into statements (an approved tool, a sanctioned script, a forbidden practice, an exception, an escalation contact) and turns each one into training examples that show the policy APPLIED to events, not just its text, so a model learns to recognise the situation the policy describes.

You then choose the knowledge you want the models to learn: the items are listed for you to edit, keep or drop, each with the model or models it goes to, and only what you keep is imported.

What stays fixed

- Each model's instructions and answer format are pinned. An adapter that breaks them is refused before it is imported.
- The same two models are tuned in place; the pipeline does not grow.
- Your documents, training files and adapters stay on this machine, and are removed with everything else when AegisAI is uninstalled.
- Every adapter is reversible: the previous one is always kept.

COMING NEXT - YOUR OWN RULES AND SIGNATURES

Personal Sigma, YARA and other rules and signatures, beside the trained models

The same ownership extends to every detection engine. The user, or an SME's security team, will add personal Sigma behavioural rules, YARA rules for files and process memory, ClamAV-format signatures and other detection content -- written in place or imported -- test them on the machine against its own telemetry before they are switched on, and keep them apart from AegisAI's own updates, so that no update ever overwrites them.

Together with the trained models, that makes AegisAI the first endpoint product whose detection logic AND whose AI are both shaped by the customer: threats defined by the organisation's own set-up and policies are detected by its own rules and understood by its own models.

your own Sigma rules | your own YARA rules | your own signatures | never overwritten by updates

The Model fine-tuning page is in the app today; its controls are switched on in the finished product.

Where the finished AegisAI goes

The finished AegisAI is a five-layer platform that delivers full extended detection and response on ordinary PC hardware with zero mandatory internet dependency. Telemetry flows up through the layers; decisions, lockdowns and policy updates flow down. Protection does not stop at detection: when a threat is confirmed, AegisAI contains it, remediates it and rolls the damage back -- automatically, or after one-click specialist approval -- and escalates anything uncertain to a human review queue. The whole detect-respond-recover loop runs on the endpoint.

BUILT AROUND ONE MACHINE AND ITS OWNER

Every layer is designed for one machine and one owner first. The platform runs on ordinary PC hardware with no GPU, uses the NPU where there is one, keeps every capability with no network at all, and grows into an SME fleet without moving a byte of telemetry outside the company. And every layer is the owner's to shape: which engines run and in what order, how far the response may go on its own, which rules and signatures are active -- the customer's own among them -- and what the two language models have been taught about the environment they protect.

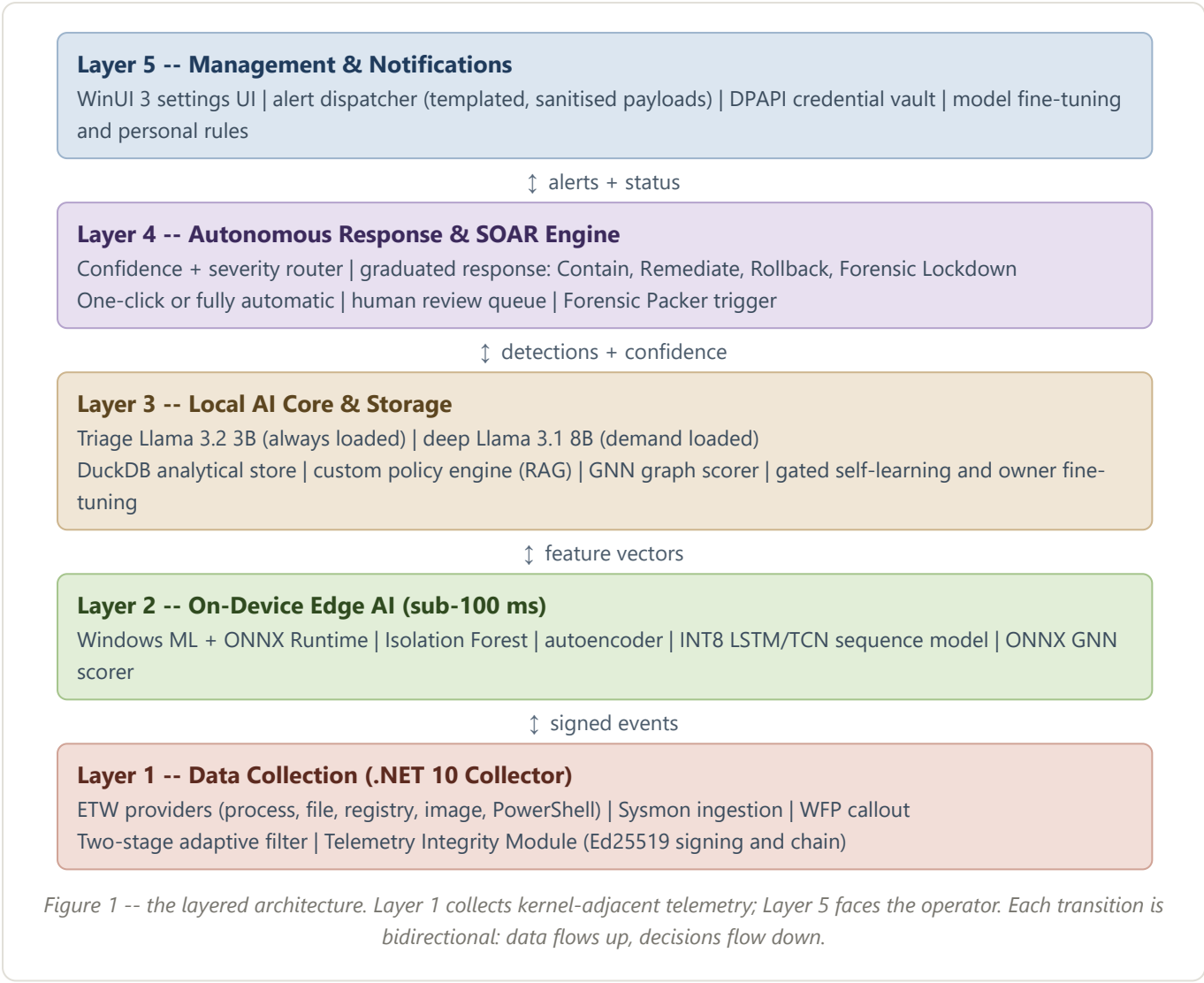


Figure 1 -- the layered architecture. Layer 1 collects kernel-adjacent telemetry; Layer 5 faces the operator. Each transition is bidirectional: data flows up, decisions flow down.

	This build	The finished product
Telemetry	ETW and Sysmon, signed and chained at the collector. AegisAI's WFP driver is built and held back until it is signed.	Adds WFP network blocking before a packet leaves, MiniFilter ransomware-entropy signals and a PID-controlled adaptive filter.
Edge AI	Two ONNX scorers: supervised behaviour and unsupervised novelty.	Isolation Forest, autoencoder, INT8 sequence model and a GNN process-graph scorer under Windows ML, on the NPU where there is one.
Language models	Llama 3.2 1B triage and a fine-tuned Llama 3.2 3B. In this prototype they explain and do not decide; in the final product they decide too, and find 0-day threats nothing else detects today.	Always-loaded Llama 3.2 3B triage at about 80 ms, Llama 3.1 8B on demand. Their verdict and confidence drive the response, and both can be fine-tuned by their owner on the owner's own policy and incidents.
Detection content	157 curated Sigma rules, YARA rules and ClamAV signatures.	The same open content, kept current by updates, plus the customer's own Sigma, YARA and other rules and signatures, which no update ever overwrites.
Response	Restorable quarantine and process suspension behind a safety floor, plus pause, close, force-kill and kill-tree controls offered only when they can work, and a Declare Safe that reverses what was done to a file.	A graduated ladder -- Contain, Remediate, Rollback, Forensic Lockdown -- one-click or automatic, per severity.
Recovery and evidence	Detection journal and an action audit, append-only except for files a user declares safe.	Surgical rollback from shadow copies and the signed change-journal; sealed, signed .aegis forensic archives.
Learning	Fixed models, shipped with the product.	A gated self-learning loop: signed specialist feedback, LoRA adapters over a frozen golden model, shadow evaluation of every high-confidence verdict -- and direct fine-tuning by the owner from the Model fine-tuning page.
Fleet	One endpoint.	AegisAI grows from single-endpoint protection into a full fleet protection tool for SMEs: an optional on-premises console for up to 250 endpoints, FIDO2-gated, over endpoint-initiated mTLS. Everything stays inside the company, with no cloud anywhere, and endpoints lose nothing when the console is unreachable.

A dual-LLM pipeline that explains today -- and will decide and suggest once trained

In this build the two Llama models only explain. That is a deliberate rule, not a limit of the design: models that have not yet been trained on real endpoint telemetry should not be allowed to override a detection, a severity or a response, and here they cannot.

IN THE FINISHED PRODUCT

In the final product the fine-tuned dual-LLM pipeline will also make decisions, based on its own explanations and analyses of what a file or a process does or may do. It will discover 0-day threats that nothing else available today can detect, and it will detect threats defined by you: by the particulars of your computer's set-up and your own policies, or, in AegisAI Local - SME Edition, by your company's security policies and practices.

The finished product takes the same pipeline further. The always-loaded triage tier processes every event the edge models promote and returns a natural-language verdict with a confidence score. When the confidence is high -- about 95% of cases -- that verdict goes straight to the response engine; when it is low, or the case is novel, the deep tier is called in for a longer, more nuanced attack storyline. The response engine then combines the AI's confidence with the incident's severity to choose the least destructive action that fits: in one-click mode the pipeline SUGGESTS -- the action is fully prepared and a specialist approves it with a single confirmation -- and in fully automatic mode it DECIDES, for the cases where waiting is riskier than acting.

Deciding is earned, and the finished product fences it in. Learning happens only through signed specialist feedback that passes a Feedback Integrity Gate, or through the owner's own fine-tuning on the Model fine-tuning page, where every adapter is tested on held-back examples before it is used; base weights are never modified, only LoRA adapters layered over a frozen golden model; on every high-confidence detection the golden and fine-tuned models are compared, and a divergence is flagged before any response is taken. A low-confidence verdict never destroys anything -- it falls back to containment and the human review queue.

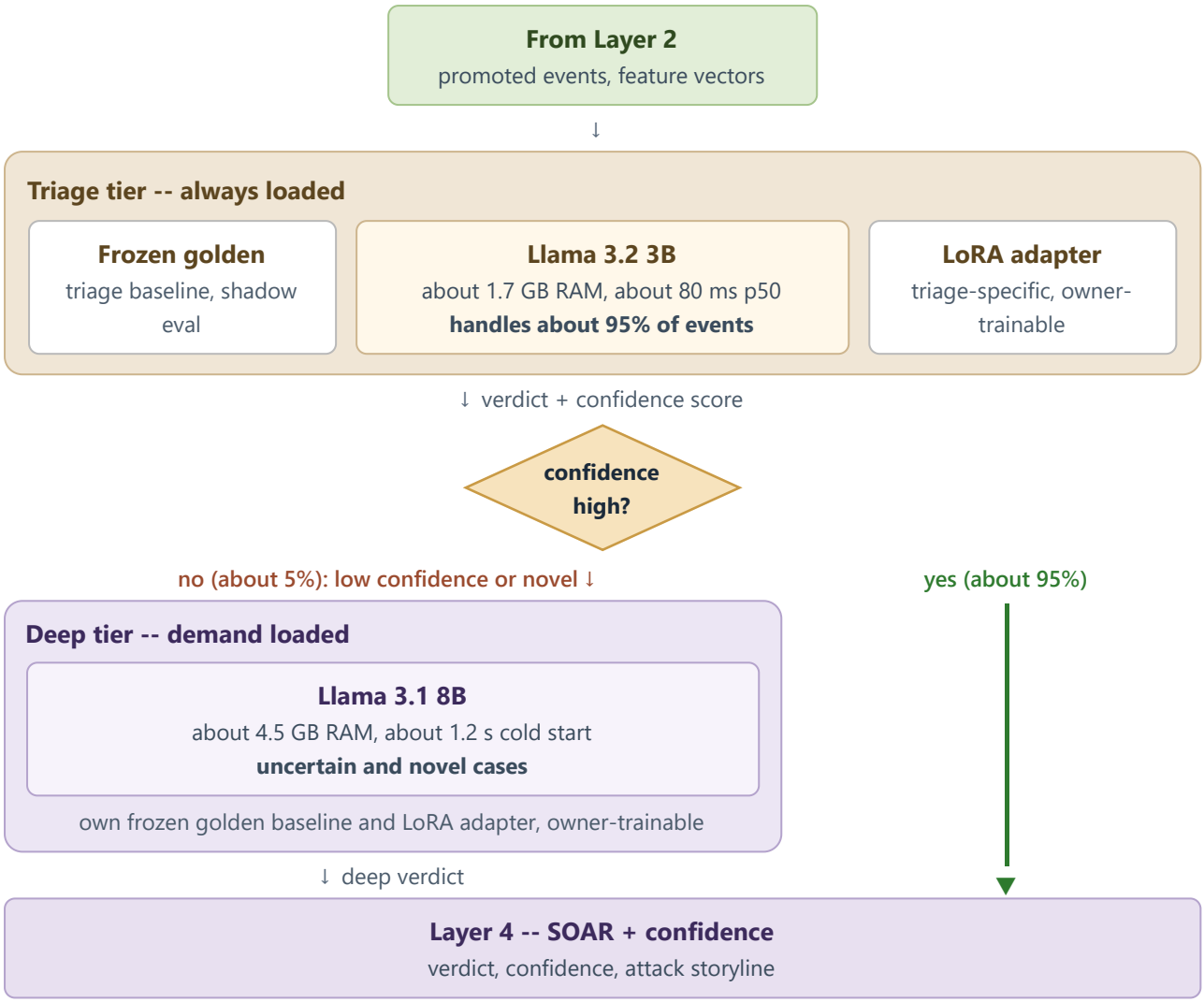


Figure 2 -- the dual-LLM tier. Every promoted event is read by the always-loaded triage model; the confident majority goes straight to the response engine and the uncertain or novel few escalate to the deep model. Each tier keeps its own LoRA adapter and frozen golden baseline for shadow evaluation.

A graduated, reversible response

The finished product replaces the all-or-nothing reaction of most endpoint tools with a ladder. A confidence-and-severity router picks the least destructive rung that fits the threat; every rung is rate-limited, written to the tamper-evident chain and reversible wherever the operation allows; and any action seals a signed evidence archive. Rollback -- the on-device equivalent of cloud storyline rollback -- is the primary recovery path against ransomware-style encryption.

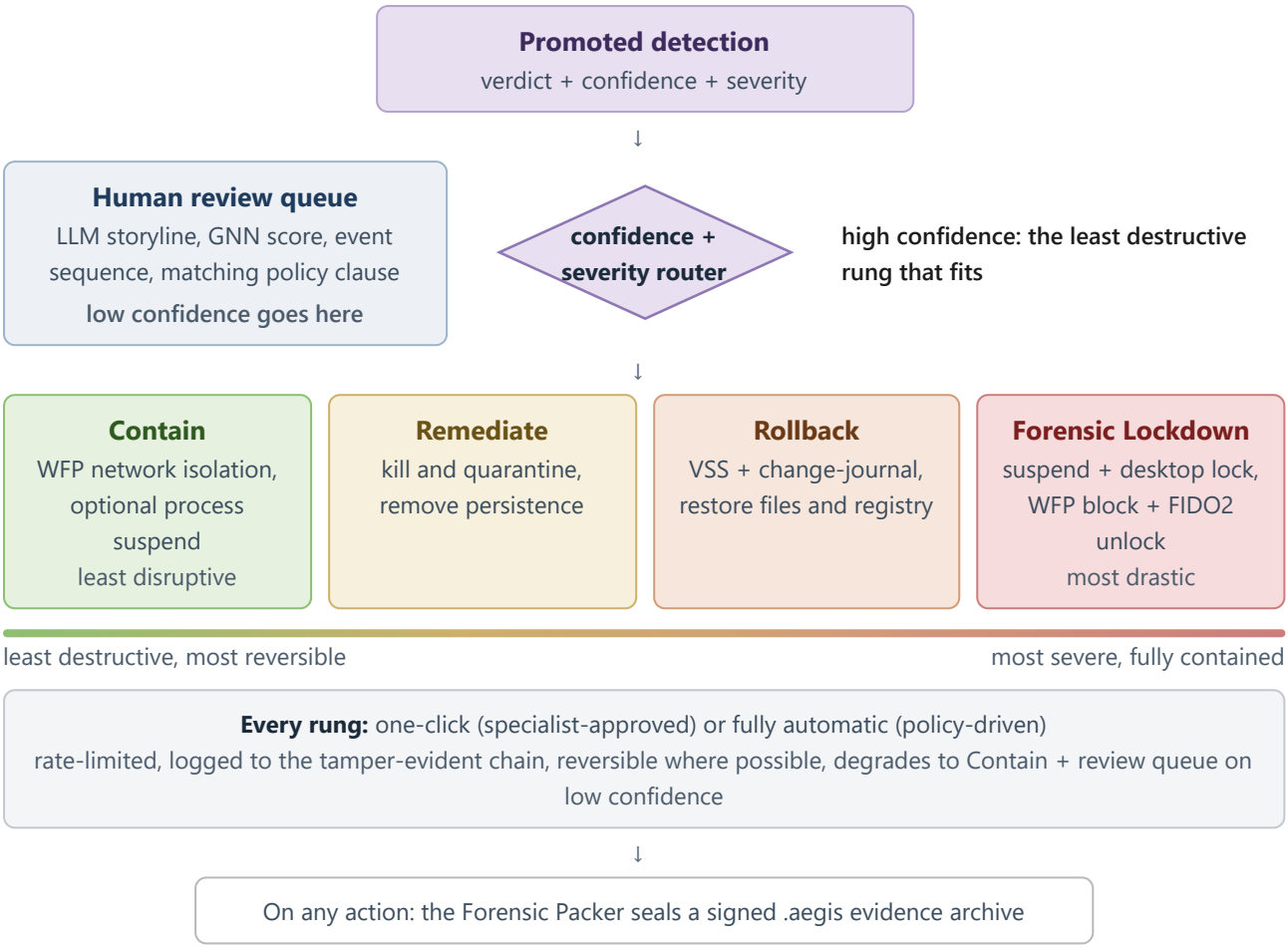


Figure 3 -- the graduated response ladder. The router selects the least destructive rung that fits the threat; low-confidence detections divert to human review. Each rung runs one-click or fully automatic, is rate-limited and logged, and degrades safely.

Together, something the competition cannot offer

AegisAI runs the whole security loop -- detect, explain, respond, recover and learn -- on the machine it protects. The leading endpoint platforms already detect on the device, but the correlation, the hunting and the generative AI that makes sense of an incident live on the vendor's platform, usually its cloud, and serve an analyst at a console. In AegisAI every layer, both language models included, sits inside the endpoint, by design.

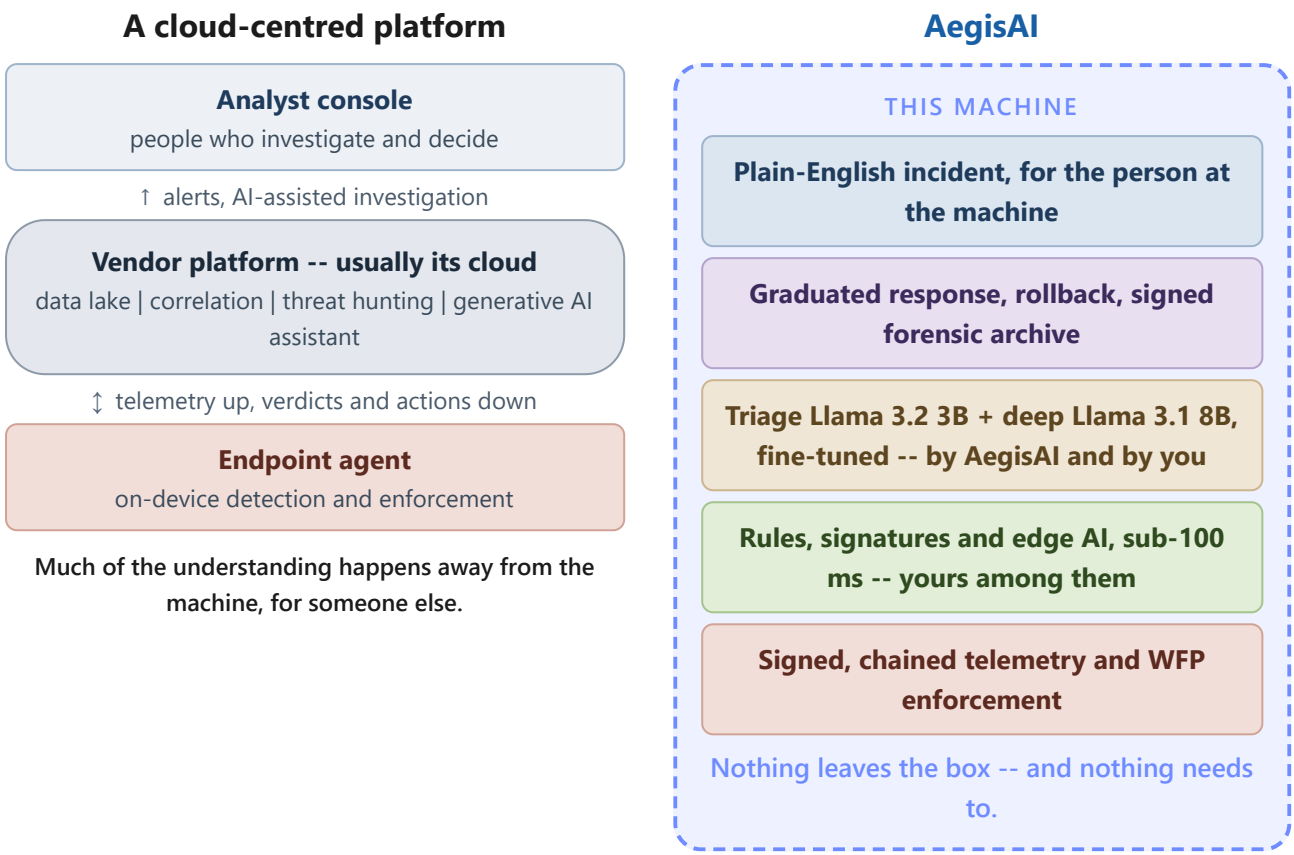


Figure 4 -- where the intelligence lives. A cloud-centred platform detects on the device and understands in its own platform; AegisAI detects, understands, responds, recovers and learns on the machine it protects.

Today's leaders were built around that platform. Their generative assistants help analysts investigate after the fact; the person at the affected machine gets an alert, not an explanation. Response is largely kill, quarantine or isolate, with shadow-copy rollback in the best of them. Their models are tuned centrally, on data pooled from all customers, so they learn the average customer rather than your machine. And their cost scales per endpoint, every year, to fund the pipeline behind them.

Sovereign by construction

Kernel-level telemetry signed and hash-chained at the source; detection, both language models, response, rollback and forensics, all on the endpoint. A machine with no network at all keeps every capability -- what air-gapped sites, regulated and privacy-bound organisations, and anyone unwilling to let endpoint data leave the box require. In AegisAI that is the design, not a separate edition.

Layered evidence, earned confidence

Deterministic detectors give verdicts you can audit line by line: SigmaHQ behavioural rules, YARA on files and in memory, and signatures. An edge-AI stack -- Isolation Forest, autoencoder, INT8 sequence model and a process-graph GNN, in under 100 ms under Windows ML, on the NPU where there is one -- catches what no rule describes. Agreement across layers is what raises severity.

Enforcement at the kernel boundary

AegisAI's own WFP callout driver blocks a connection before its first packet leaves the machine, and MiniFilter entropy signals catch ransomware-style mass encryption as it starts.

Safe autonomy

A confidence-and-severity router picks the least destructive rung that fits: Contain, Remediate, Rollback or Forensic Lockdown. Every action is rate-limited, written to the tamper-evident chain and reversible where the operation allows; rollback restores files and registry from shadow copies and the signed change journal; every action seals a signed forensic archive. A low-confidence verdict never destroys anything.

Explained where it happened

Every incident arrives with a written storyline mapped to MITRE ATT&CK, with the rules, their authors and the evidence behind it -- tier-1 analysis without a SOC, delivered to the person who owns the machine, not only to an analyst at a console.

It learns your machine, privately

Signed specialist feedback trains small LoRA adapters over a frozen golden model, and shadow evaluation compares the two on every high-confidence verdict, flagging any divergence before a response is taken. AegisAI adapts to what is normal on this machine, not on the average customer's, and no data leaves to do it. The owner can drive that learning directly, from the Model fine-tuning page.

Ownership the cloud cannot match

A perpetual licence with no compulsory retainer. An optional retainer, from USD 5 per machine per month, brings new rules for every engine, definitions, signatures and better fine-tuned Llama models. With no per-endpoint cloud pipeline to fund, that is a fraction of the cost of cloud XDR, fully air-gapped installs included. An update never overwrites the customer's own rules, Sigma or any other, and never removes LoRA fine-tuning done by the customer or, in an SME deployment, by its own cybersecurity team.

The finished AegisAI beside SentinelOne Singularity

SentinelOne is among the strongest endpoint platforms on the market. The first two rows are mostly equivalent; on every other row -- where an incident is understood, who it is explained to, how the product learns, what it costs to own -- AegisAI is ahead.

	SentinelOne Singularity	AegisAI -- the finished product
On-device detection mostly equivalent	Static and behavioural AI engines on the agent. Keeps protecting offline.	Sigma rules, YARA, signatures and four edge models on the device. Keeps protecting offline.
Ransomware recovery mostly equivalent	Rollback from Windows shadow copies, including registry keys and scheduled tasks.	Surgical rollback from shadow copies and a signed change journal, files and registry.
Incident explanation AegisAI is better	Purple AI: a generative assistant for analysts investigating in the SentinelOne console.	Every incident explained automatically, on the endpoint, by two fine-tuned local Llama models -- for the person at the machine, with no console or analyst required. **Why:** it is written on the machine, for its owner, not for an analyst in a vendor console.
Where the AI reasoning runs AegisAI is better	Detection on the agent. Correlation, hunting and Purple AI on the SentinelOne platform -- cloud by default, with on-premise editions for regulated and air-gapped sites.	All of it on the endpoint, both language models included. **Why:** nothing has to leave the machine, so air-gapped operation is the default, not a separate edition.
Detection content AegisAI is better	Proprietary engines and models.	Open, auditable content -- SigmaHQ rules with their authors credited, YARA and ClamAV -- beside the local models, plus the customer's own rules and signatures. **Why:** the rules are open and auditable, and each author's name travels with the verdict.
Response AegisAI is better	Kill, quarantine, remediate, network quarantine and rollback, per policy.	A four-rung ladder chosen by a confidence-and-severity router. Low confidence never destroys anything, and every action seals a signed evidence archive. **Why:** it takes the least destructive step that fits, and a low-confidence verdict cannot destroy anything.
Evidence integrity AegisAI is better	Anti-tamper protection of the agent.	Telemetry Ed25519-signed and hash-chained at the source. Sealed, signed .aegis forensic archives. **Why:** the record is signed and hash-chained as it is collected, not only shielded from tampering afterwards.
Learning AegisAI is better	Models trained and updated centrally by the vendor.	Per-machine LoRA adapters from signed specialist feedback, shadow-evaluated against a frozen golden model -- and fine-tuning by the owner, on the owner's own policy and incidents. **Why:** it adapts to this machine, not to the average customer.
Value and ownership AegisAI is better	Per-endpoint annual subscription, in tiers.	A perpetual licence, no compulsory retainer; an optional retainer from USD 5 per machine per month for new rules, definitions, signatures and fine-tuned models. **Why:** no cloud pipeline to fund. Updates sit beside what the customer owns and never overwrite its own rules or remove its own fine-tuning.

SentinelOne, Singularity and Purple AI are trademarks of SentinelOne, Inc.; its column reflects public information as of November 2026.

Why properly fine-tuning the two Llamas is the multiplier

A stock language model knows English, not endpoint telemetry. Sysmon event IDs, parent-child process chains, obfuscated command lines and the exact boundaries between ATT&CK techniques are a specialist grammar, and a general model reads them the way a well-read layperson reads an X-ray. Fine-tuning on real, labelled telemetry turns the Llama 3.2 3B triage model and the Llama 3.1 8B deep model into specialists -- and that changes five things.

	A stock general model	A fine-tuned AegisAI specialist
Reads telemetry	like a well-read layperson	**like a trained analyst**
Grounding	fills gaps with plausible stories	**cites what the rule asserts; says "not observed" where the evidence stops**
Confidence	a number with nothing behind it	**calibrated on labelled outcomes -- safe to act on**
Every prompt carries	instructions, format, grounding rules and the evidence	**the evidence alone, in a compact encoding**
Output	varies from run to run	**the same structured verdict for the same evidence**
Size it needs	larger, to compensate	**small enough for an ordinary PC**

Figure 5 -- what fine-tuning changes. The same base model, before and after it is trained on real, labelled endpoint telemetry.

1 Grounding: it says what the evidence shows, and nothing more

An untuned model fills gaps with plausible stories: a sibling technique instead of the right one, a command-and-control beacon inferred from a single encoded command. Trained on windows labelled against their rules and techniques, the model learns to cite the technique the rule actually asserts and to say "not observed" where the evidence stops. Whole techniques are held out of training, so this is measured on attacks the model has never seen.

2 Calibrated confidence: the key that unlocks autonomous response

In the finished product the triage tier sends its confident majority, about 95% of cases, straight to the response engine. That is safe only if "95% confident" means right 95% of the time, and a stock model's stated confidence is calibrated against nothing. Fine-tuning on labelled outcomes is what makes the number mean something -- it is what moves AegisAI from explaining to acting.

3 Speed: what the model knows, it does not have to read

On a CPU, reading the prompt dominates the time to an explanation -- about 94% of it, measured on this prototype's deep tier. A general model has to be told, on every call, the instructions, the output format and the grounding rules; a fine-tuned model carries them in its weights, so the prompt shrinks to the evidence alone, in a compact encoding the model was trained to read. Every token removed is time saved, almost linearly, and a triage tier trained to answer in a fixed structured form needs only a few dozen output tokens.

4 Consistency: the same structure, every time

A tuned model returns the same structured verdict -- technique, severity, confidence, storyline -- for the same evidence. The two tiers can then be measured against each other on every incident, and AegisAI already records where they disagree: quality becomes a number the product watches, not an impression.

5 Small and specialised beats large and general

Published results repeatedly show small models tuned to a narrow, well-defined task matching much larger general models on that task. That is what lets AegisAI's reasoning run on an ordinary PC, with no GPU and no data centre. A bigger stock model would cost memory and latency, and would still be uncalibrated on telemetry.

Why both tiers need it -- for different reasons

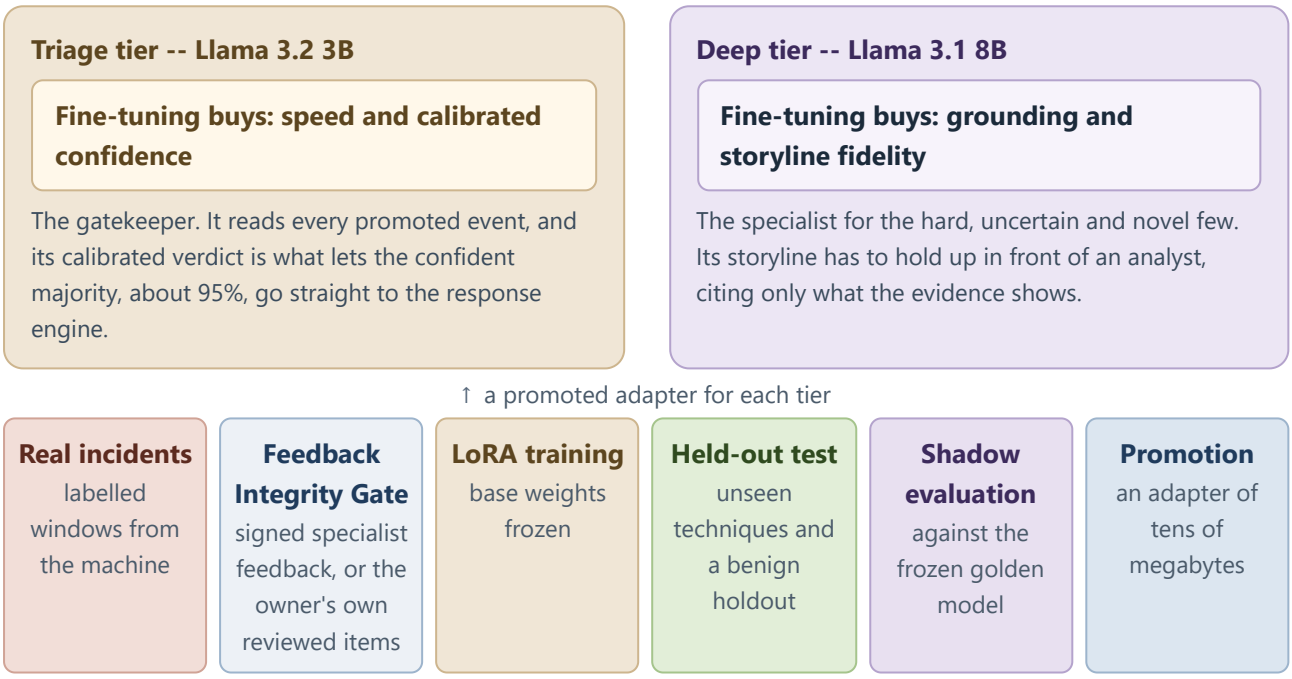


Figure 6 -- two specialists, one gated learning loop. Each tier keeps its own adapter and frozen golden baseline; an adapter is promoted only after it has passed held-out testing and shadow evaluation.

WHY ONLY A LOCAL PRODUCT CAN DO THIS

One global model for every customer -- or a specialist for every machine.

A cloud vendor tunes one global model on data pooled from all its customers. AegisAI tunes a small adapter -- tens of megabytes against gigabytes of base model -- for the machine it protects, through a gated, signed loop, without a byte of telemetry leaving it. The same adapter can reach an air-gapped site on a USB stick.

Why AegisAI can disrupt endpoint protection

Endpoint protection is dominated by a handful of large platforms that all made the same bet: send every endpoint's telemetry to the vendor's cloud, understand it there, sell generative AI to the customer's analysts as a premium layer, and charge per endpoint, every year, for as long as the customer stays. It is a strong model for a large enterprise with a security operations centre. For almost everyone else -- small and medium businesses with no SOC, regulated and air-gapped sites, privacy-bound organisations, and individuals -- it means paying for an analyst console nobody watches, for a data pipeline the customer would rather not have, and for AI the customer can neither see, nor shape, nor own.

AegisAI makes the opposite bet, and every part of its architecture follows from it.

1 The AI runs on the machine it protects -- nobody else does this

Two Llama language models read and explain every incident on the endpoint itself, on an ordinary CPU, with no GPU and no network. The established AI products put their language models in the vendor's cloud and point them at an analyst; on-device AI in today's market is limited to classic machine-learning detection models and, in consumer products, narrow NPU features such as deepfake detection. A per-incident language-model analyst inside the endpoint is new to the category.

2 The owner takes part in training the AI -- a first for the market

AegisAI ships both models extensively fine-tuned and keeps them current with regular updates. With the Model fine-tuning page the end user, or an SME's own security team, adds its own knowledge on top, as and when it needs to: upload a policy, review what AegisAI extracted from it, train a LoRA adapter on your own hardware, test it against held-back examples, switch it on, and go back with one choice. Every established vendor trains its models centrally and ships the same model to every customer; none of them lets the customer take part. With AegisAI the AI becomes partly the customer's own.

3 Threats defined by you, not by the average customer

A cloud model knows what is malicious across all customers; it cannot know that your accounting PC must never run PowerShell, or that the 2 a.m. script is your backup. AegisAI's tuned models, and in future its personal Sigma, YARA and signature rules, detect what is wrong for THIS machine and THIS company -- insider misuse, policy breaches and unusual-for-us behaviour that no global model and no generic rule set can define.

4 Not one-size-fits-all, at any layer

Choose which engines scan and in what order, the dual-LLM pipeline among them; decide how far the response may go on its own, from one-click approval to fully automatic per severity; add your own detection content beside AegisAI's; keep the models AegisAI ships or teach them your environment. The same product fits a single professional's laptop, a 20-person office and a 250-endpoint SME fleet with its own console, without a byte leaving the company.

5 Price: a structural advantage, not a discount

A cloud XDR subscription funds a per-endpoint data pipeline, a data lake, retention and GPU inference for its AI assistant, every year. AegisAI has none of those costs: the language models run on the customer's own CPU, nothing is stored by the vendor, and the cost of protecting one more endpoint is close to zero. That lets AegisAI sell a perpetual licence with no compulsory retainer -- the optional retainer for updates starts at USD 5 per machine per month -- and still be profitable, at a small fraction of what the market charges today -- while competitors cannot follow without dismantling the cloud model their revenue depends on.

6 Sovereign and private by default

Data residency, GDPR and sector regulation are pushing buyers away from sending endpoint data abroad. Where incumbents answer with separate on-premise or air-gapped editions, AegisAI is air-gapped by construction: full capability with no network at all, no data retention, and models and adapters that reach an isolated site on a USB stick.

7 Open and auditable

Detection rests on open, community-reviewed content -- SigmaHQ rules with every author credited, YARA and ClamAV -- beside the local models. Every verdict can be traced to the rule, the engine and the evidence behind it, and every action is signed, chained and reversible. That is what auditors, insurers and regulated buyers ask for and what black-box proprietary engines cannot show.

8 Riding the hardware wave

Every new business PC ships with an NPU and enough memory to run small language models locally, and small models tuned to a narrow task keep closing on large general ones. The cost of on-device AI falls every year; the cost of cloud AI per endpoint does not. AegisAI is built for the hardware buyers are already purchasing.

What the market charges today

List or reported prices per endpoint, before the generative-AI layer where that is sold separately. For 100 endpoints over three years, SentinelOne Singularity Complete at list price comes to $53,997 and CrowdStrike Falcon Enterprise to $55,497, every three years, for as long as the subscription runs. AegisAI's optional retainer for the same 100 machines, at USD 5 per machine per month, comes to $18,000 over three years -- and protection continues without it.

Product	Price (list or reported)	Generative AI
SentinelOne Singularity	Complete $179.99, Commercial $229.99 per endpoint per year (list) [1]	Purple AI, in "includes Purple AI" packages [1]
CrowdStrike Falcon	Go $59.99, Pro $99.99, Enterprise $184.99 per device per year (reported list) [2]	Charlotte AI, sold separately (reported per analyst seat or per action) [2]
Microsoft Defender	Defender for Business $3 per user per month, up to five devices [3]	Security Copilot: included compute with Microsoft 365 E5; otherwise $4 per provisioned compute unit per hour [4]
Palo Alto Cortex XDR	Pro about $81-120 per endpoint per year (third-party estimates) [5]	Cortex AgentiX, quote-based [5]
Huntress Managed EDR	$8.99 per endpoint per month (50-99 endpoints), SOC included [6]	-- (human SOC)
AegisAI	Perpetual licence, no compulsory retainer. Optional retainer from USD 5 per machine per month for new rules, definitions, signatures and fine-tuning files, added beside the customer's own	Both language models included, on the endpoint, trainable by the customer

Who buys first

SMEs without a security team

They need the explanation and the next step, not a console. AegisAI gives tier-1 analysis on every machine and, in the SME Edition, a company-internal fleet console -- at a price an office can afford.

Regulated and air-gapped sites

Defence, government, healthcare, finance, critical infrastructure and industrial networks, where endpoint data may not leave the site and cloud AI is not an option at all.

Professionals and privacy-conscious users

Lawyers, doctors, accountants, journalists and anyone who wants enterprise-grade protection without sending their machine's activity to a vendor.

MSPs and integrators

A product they can tune per client -- each client's policy in its own adapter, its own rules beside AegisAI's -- with no per-endpoint cloud bill eating their margin.

The market today: who comes closest

We surveyed the leading AI-driven endpoint platforms and the open, self-hosted alternatives in November 2026. Several run machine learning on the endpoint for detection; several offer a generative assistant; a few can be self-hosted with a local language model on a server. We found none that runs language-model reasoning on the protected machine itself, explains each incident to that machine's owner, and lets the owner train the AI on their own policy. AegisAI is the only product that does all three.

● *yes* ◐ *partly* ○ *no, or not found in the vendor's public material.*

Product	Language-model reasoning on the protected machine	Explains each incident to the machine's owner	Full capability with no network	Owner takes part in training the AI	Customer's own detection rules	Open, auditable detection content
AegisAI (finished product)	● two Llama models, on the CPU	● every incident	● air-gapped by design	● Model fine-tuning page, LoRA	● Sigma, YARA, signatures	● SigmaHQ, YARA, ClamAV
SentinelOne Singularity + Purple AI	○ Purple AI on the platform	○ analyst console	◐ agent protects offline; on-prem editions since Mar 2026 [7]	○ vendor-trained	● custom rules	○ proprietary
CrowdStrike Falcon + Charlotte AI	○ cloud platform	○ analyst console	◐ agent protects offline	○ vendor-trained	● custom IOA rules	○ proprietary
Microsoft Defender + Security Copilot	○ Microsoft cloud	○ analyst portal	◐ AV protects offline	○ vendor-trained	● custom detections	○ proprietary
Elastic Security (Elastic Defend + AI Assistant)	◐ self-hosted LLM on a server [8]	○ analyst in Kibana	◐ with self-managed servers [8]	◐ bring your own model; no tuning workflow	● custom rules	● open rule repository
Wazuh (open source)	◐ local LLM threat hunting on the server [9]	○ analyst dashboard	◐ self-hosted servers	◐ bring your own model; no tuning workflow	● own rules; Sigma in 5.0 [10]	● open source
Arctic Wolf Aurora (Cylance)	○ classic ML, no language model	○ managed SOC	◐ ML built for limited connectivity [11]	○ vendor-trained	◐ policy settings	○ proprietary
Norton (Gen Digital), consumer	◐ NPU models for deepfake and scam checks only [12]	○ scam assistant, not incidents	◐ core AV offline	○ vendor-trained	○ --	○ proprietary

Closest in concept, and where they stop

Elastic Security and Wazuh come closest on openness and data control: both can keep data on the customer's own servers and both can connect a locally hosted language model. But the model runs on a server the customer has to build and run, it serves an analyst at a dashboard, and tuning it is the customer's own machine-learning project. Neither puts a language model on the protected machine or offers a guided, tested, reversible fine-tuning workflow.

SentinelOne is the strongest commercial platform and, since March 2026, offers on-premise and air-gapped editions for regulated sites. Its generative AI, Purple AI, remains an assistant for analysts on the SentinelOne platform, and its models are trained by SentinelOne for all customers.

CrowdStrike and Microsoft have the largest generative-AI investments in the category -- Charlotte AI and Security Copilot -- and both run them in their own clouds, metered per analyst seat, per action or per compute unit, on top of the endpoint subscription.

Arctic Wolf Aurora (Cylance) and Norton show that on-device AI works: Cylance pioneered offline machine-learning prevention, and Norton runs NPU models for deepfake and scam detection on AI PCs. Neither runs a language model that reads, explains and acts on endpoint incidents, and neither lets the customer train it.

THE GAP AEGISAI FILLS

Local reasoning, explained to the owner, trained with the owner, on open rules, at a perpetual-licence price. Each of the products above has one or two of these. AegisAI is designed around all of them.

Sources (public information, checked November 2026; prices are list or reported and vary by contract). Product names are trademarks of their owners.

[1] SentinelOne pricing and platform packages: sentinelone.com/pricing, sentinelone.com/platform-packages
[2] CrowdStrike Falcon and Charlotte AI pricing guides: casrai.org/guides/crowdstrike-pricing, softwarecontractnegotiation.pages.dev/blog/crowdstrike-charlotte-ai
[3] Microsoft Defender for Business: microsoft.com/security/business/endpoint-security/microsoft-defender-business-b
[4] Security Copilot inclusion in Microsoft 365 E5: learn.microsoft.com/copilot/security/security-copilot-inclusion, samexpert.com/security-copilot-licensing-guide
[5] Cortex XDR estimates and Cortex AgentiX: costbench.com/software/cybersecurity/palo-alto-networks, investors.paloaltonetworks.com
[6] Huntress pricing: underdefense.com/blog/huntress-pricing-guide
[7] SentinelOne on-premise and air-gapped editions, 23 March 2026: businesswire.com/news/home/20260323813450
[8] Elastic, self-managed LLMs: elastic.co/docs/explore-analyze/ai-features/llm-guides/local-llms-overview
[9] Wazuh local-LLM threat hunting: securityinfowatch.com (Wazuh press release)
[10] Wazuh Sigma rules: documentation.wazuh.com/5.0-rc/user-manual/wazuh-indexer/wazuh-sigma-rules.html
[11] Arctic Wolf acquisition of Cylance: itpro.com, techzine.eu
[12] Norton on-device deepfake protection: newsroom.gendigital.com, gendigital.com

The developer



Fotios Basagiannis

IT Systems Architect, Lead Developer and Cybersecurity Engineer

Blumenau, Brazil | author of the AegisAI specification

More than two decades as a hands-on systems architect and lead developer across the private and public sectors, including six years of cybersecurity engineering at the European Commission's Joint Research Centre. AegisAI is built the way that work taught: deterministic detection first, local AI for explanation, and a design that refuses to claim protection it has not established.

Led ERCA for the European Commission
Led the development and first successful production deployment of ERCA, the European Root Certification Authority of the EU digital tachograph system -- a project of major significance that is still part of the EU's IT infrastructure today, increasing safety on Europe's roads.

Six years of cybersecurity R&D at the Joint Research Centre
Cybersecurity Engineer at the European Commission's JRC in Ispra, Italy, from 2003 to 2009, working on EEJ-Net, PIPS, ERCA and MEDI.

Trusted agents inside hardware security modules
Designed from scratch and led a proof of concept of a secure peer-to-peer network of trusted agents embedded in HSMs, for the exchange of sensitive information between EU governments.

Top-rated consultant in AI-augmented systems since 2010
AI-augmented full-stack systems, Windows internals and background services, low-level C and C++ automation, cryptography and HSMs, and multimodal AI document analysis -- delivered to clients worldwide.

Educated in the UK and the US
MSc Computer Science, Heriot-Watt University, Edinburgh (2002). BSc Computer Science, cum laude, The City College of New York (2000).