CAELARA LABS

Enforceable Boundaries for Intelligent Systems: The Caelara Labs Architecture for Pre-Execution Security, Governed Routing, and Verifiable Receipts

Caelara Labs Research & Engineering · November 2026 · https://caelaralabs.com

Whitepaper · https://caelaralabs.com/whitepaper/

Contents

Abstract

1. Executive thesis

2. The non-custodial governance principle

3. Architecture: six interconnected product families

4. Heuristic screening boundaries

5. Threat model

6. Public telemetry

7. Dual-use and sovereign edge readiness

8. Conclusion

Abstract

Intelligent systems now read untrusted text, follow links and take actions on behalf of people and businesses. Most of the risk sits in the moment before an action runs: an instruction hidden in a web page, a leaked credential in a reply, a link that leads somewhere it should not. Caelara Labs builds the layer that sits in that moment. It screens what goes in and what comes out, routes calls through one governed gateway, and issues a receipt for every verdict that anyone can check later. This paper describes that architecture, the principle that keeps it non-custodial, the boundaries of what heuristic screening can promise, and the public telemetry that lets customers judge it for themselves.

1. Executive thesis

> ### Interact. Decide. Act. With enforceable boundaries.

Caelara Labs is an independent technology company working across research, engineering, infrastructure and product development. Its products form one integrated ecosystem whose purpose is simple to state: an intelligent system should be able to interact, decide and act, and every one of those steps should be bounded by a policy that is enforced before execution and provable afterwards.

Four practical needs drive adoption:

1. **Stop risky inputs and links before they execute.** Prompts, model output, links and requested actions are screened at the edge, before an application or agent acts on them.
2. **Prove what was decided without storing what was said.** Every paid screening returns a tamper-evident receipt built from SHA-256 hashes. The receipt can be checked by anyone at `caelaralabs.com/verify/<sha256>`, and no request text is kept.
3. **Replace a patchwork of vendors with one gateway.** One key and one MCP-compatible gateway reach Caelara Shield and ten API suites, instead of a separate contract, credential and integration for each capability.
4. **Adopt progressively.** A team can start with keyless receipt verification and the free in-browser check page, move to a free trial key, then to one-time prepaid EGU packs or per-call payment, without changing integration and without a subscription.

2. The non-custodial governance principle

> **Software that decides, never custody.**

Caelara Labs builds software for policy evaluation, pre-execution screening and verifiable audit receipts. It never holds private keys, never takes custody of funds, and never executes or transmits a payment on anyone's behalf. A Caelara Labs response is a software verdict and a receipt; acting on that verdict remains with the customer's own systems.

This principle is structural, not a policy statement layered on top. The screening services accept text, links and action descriptions and return verdicts. They do not hold credentials for the systems they protect and do not build anything for another party to sign. Where earlier experimental features did build actions for external signing, they were retired and now answer `410 Gone`.

3. Architecture: six interconnected product families

The ecosystem is presented as six product families. Each is useful on its own; together they form one path from a first screening to a governed, metered production deployment.

3.1 Caelara Shield

Pre-execution security. Caelara Shield screens four surfaces:

SURFACE	WHAT IT CHECKS	PRICE PER CALL
Prompt	Instructions hidden in user or retrieved text, attempts to extract system instructions, spoofed system or administrator markers, overrides written in other languages, and payloads hidden in encodings	$0.015
Output	Model replies for leaked credentials and personal data, exfiltration links and active content	$0.015
Proxy	Untrusted URLs fetched through a guarded proxy that refuses private, local and disguised network targets, with the page scanned on the way back	$0.025
Action authorisation	An action an agent is about to take, checked against the customer's own rules (allowed domains, per-action and daily limits, an approval threshold); the answer is allow, deny or require approval, with a one-tap approval link	$0.005

An optional boot clearance ($0.005) screens an agent's system prompt and tool descriptions once at start-up.

Every paid screening returns an `audit_receipt`: the surface, the verdict, a risk level, a timestamp, a digest of the exact request and a hash of the canonical record. Only that canonical record of hashes, verdict and time is stored, for 400 days. Receipts are checked with a keyless request to `/v1/shield/receipt/<hash>` or in a browser at `caelaralabs.com/verify/<sha256>`, which shows the verdict as a card and returns the exact JSON to programmatic clients.

3.2 Intelligence APIs

Ten routed API suites behind one governed gateway: web and company intelligence, social analytics, agent safety, compliance and trust, commerce intelligence, search and readiness audits, document transforms, locale and time, and media generation. Each call is either atomic ($0.005) or composite ($0.025). Drop-in compatibility routes let a team change a base URL and keep its existing request shape. Every suite publishes an OpenAPI specification, an MCP server and an agent card.

3.3 Developer Ecosystem

Builders reach the platform through the tools they already use: `@caelaralabs/mcp` and `@caelaralabs/shield-mcp` on npm, servers listed in the Official MCP Registry under `com.caelaralabs/*`, Agent2Agent endpoints, open-source starter kits under the MIT licence, and machine-readable discovery files (`/llms.txt`, `/.well-known/agent.json`, `/openapi.json`). The developer workbench at `caelaralabs.com/workbench/` runs live probes against production.

3.4 Agentic Commerce

Applications and autonomous agents pay for exactly what they use. Access is by API key (`X-Caelara-Key`) on a published price list:

PATH	TERMS
Free trial	50 calls over 7 days, no card
Free developer key	10,000 EGU a day
Prepaid EGU packs	$5, $25, $100 or $500 one time at $0.0002 per EGU, with 5% and 10% bonus EGU on $100 and $500; atomic 1 EGU, composite 5 EGU
Per call, no key	$0.005 atomic, $0.025 composite, settled only on a served answer

Quotas are enforced at the gateway. A request without credentials, or beyond its quota, receives HTTP 402 with machine-readable payment terms and a direct link to upgrade or start a trial, so every refusal is also a way to proceed. Prices are published before a call is made, and nothing charges a person or an agent without their action.

3.5 Autonomous Operations

Caelara Labs runs its own systems the way it asks customers to run theirs. Health monitoring, verification and publishing operate inside the same policy gates the platform sells, and their results are published rather than asserted: a public status summary and measured latency benchmarks per suite.

3.6 Consumer Applications

The same screening reaches people directly. The 1-Tap web app at `caelaralabs.com/app/` screens a link or a pasted message with one switch and keeps numbered receipts. The Telegram assistant `@CaelaraLabsBot` screens what is pasted into a chat and opens the same app inside Telegram. The check page at `caelaralabs.com/check/` runs a screening engine in the browser with no sign-in.

---

4. Heuristic screening boundaries

Screening is heuristic. A clean result is not proof that content is safe, and a block is not proof of intent. Caelara Labs states these boundaries plainly so that customers can place the layer correctly in their own defences:

- **One message at a time.** Shield screens each request on its own. An attack split across several turns of a conversation can pass each individual check; this is the known category of miss in in-house testing.
- **In-house evidence, not an independent audit.** Detection results come from corpora written and run by Caelara Labs. Against 240 constructed attack variants the prompt screen caught 237, and it flagged 1 of 55 tricky benign inputs. Six held-out attack prompts, committed before the rules were written, were all caught with no false positives on their six benign counterparts. These figures describe internal testing and are not a guarantee.
- **Not a replacement for authorisation.** A verdict says whether content looks unsafe. It does not decide whether a user is entitled to an action; that decision stays with the customer's own access controls.
- **Bounded inputs.** Proxy fetches stop at a fixed body size and time budget and return a partial result rather than waiting indefinitely.

5. Threat model

In scope. Untrusted text that tries to redirect an application or agent (direct and indirect prompt injection, instruction extraction, spoofed authority markers, multilingual and encoded overrides); replies that leak credentials, personal data or active content; links and fetches aimed at private or disguised network targets; and agent actions that fall outside the customer's own rules or exceed their limits.

Integrity of receipts. A receipt binds a verdict, a risk level, a time and a digest of the request into one hash. Changing any field changes the hash, and the verification endpoint re-hashes the stored record before it reports a receipt as valid. A receipt proves that Caelara Labs issued that verdict for that request digest at that time. It does not prove that the content was safe, and receipts are not anchored on a public ledger.

Privacy. Request text, model output and fetched pages are not stored. Trial sign-up emails are stored as keyed hashes. The check page screens on the device, in the browser.

Out of scope. Multi-turn attacks assembled across separate requests, compromise of the customer's own infrastructure or keys, and attacks that have not yet been seen in any form. Customers should combine screening with least-privilege tool access, human approval for consequential actions, and their own logging.

6. Public telemetry

Trust is easier to extend when the evidence is public:

- **Telemetry** at `caelaralabs.com/telemetry/` explains what is measured and links the live sources.
- **Status** at `status.caelaralabs.com` reports the health of every public host and a set of invariants checked on each sweep, such as retired routes staying retired and receipt errors answering correctly.
- **Benchmarks** at `benchmarks.caelaralabs.com` publish measured latency per suite. No competitor figure is published unless it can be cited.

The availability target is 99.98 percent. The measured figure is whatever the status page shows, and Caelara Labs does not quote a number it has not measured.

7. Dual-use and sovereign edge readiness

The primitives that protect a consumer app also hold where connectivity, custody and data residency are constrained: disconnected, intermittent and limited-bandwidth (DDIL) settings, and sovereign deployments that must prove what happened without calling home.

- **Offline-verifiable receipts.** Every pass and receipt is signed with Ed25519 and verifies against the published key set at `/.well-known/jwks.json` with no network call. Receipts chain in sequence, each naming the hash of the one before it, so the record of decisions stays provable through an outage and reconciles when the link returns.
- **Zero-data-retention policy gates.** Pre-execution checks answer allow, review or block over a digest of the action; the payload itself never has to leave the caller, and no request text is stored. A review verdict can escalate to a person, whose approval returns as a single-use signed pass.
- **Air-gapped deterministic sandboxes.** Code runs with no network, no filesystem and no environment, a clock pinned to the epoch, seeded randomness and fixed memory and time budgets, so the same request always yields the same signed result.
- **Scoped sub-agent delegation.** A key can mint sub-keys limited to named route families, with an expiry and a hard EGU spend cap; unspent units return when a sub-key expires or is revoked.

These are properties of the software, not certifications: no accreditation or authorization to operate is claimed.

---

8. Conclusion

Intelligent systems need boundaries that are enforced before execution and provable afterwards, delivered without taking custody of anything they protect. Caelara Labs provides that layer as one ecosystem: screening that states its limits, a gateway that replaces fragmented vendors, receipts anyone can verify, and a path from a free first check to pay-as-you-go use. Start at caelaralabs.com/app/ or caelaralabs.com/workbench/.

Powered by Caelara Labs · ops@caelaralabs.com

Powered by Caelara Labs · © 2026 Caelara Labs · ops@caelaralabs.com · https://caelaralabs.com