



AI-Driven Cybersecurity for Financial Service Providers

Reasoning You Can Audit: LLM Agents in a Regulated SOC

George Fatouros (INNOV-ACTS)

2nd ECSCI Cluster Webinar · 21 September 2026



CyberAID has received funding from the DIGITAL EUROPE Programme under grant agreement No. 101249596

What CyberAid Is

Agentic AI for financial critical infrastructure — coordinating the tools an institution already runs, and adding detection where the evidence is language or code.

BLUEPRINTS

reference architecture

PROACTIVE

vulnerability management

MONITOR

real-time detection

REPORT

incident response

Four pilots in live financial environments

Private banking & trading

Payment service providers

Retail banking

High-frequency trading

Where the LLM Actually Sits

Deterministic instruments produce the evidence. The agents read all of it — and detect for themselves only where the evidence is language or code.

INSTRUMENTS deterministic, fast, no model in the loop

eBPF kernel and network telemetry

below the application layer

SIEM / XDR

alerts and event history

Transaction graph analytics

structuring, layering, coordinated accounts

Cross-institution shared memory

patterns and context, never raw data



governed adapters · scoped retrieval

LLM AGENTS two jobs

Read from instruments/tools

Telemetry, alerts, graph results and shared patterns become evidence to correlate, corroborate and explain.

Detect where the evidence is language or code

Email threat detection · source and CI/CD analysis · log narratives.

Detection Is Not a Decision

Whichever instrument fires, the next four hours are the same problem — and they run against a statutory deadline.

classify · corroborate · decide · document

today, mostly by hand



Both DORA and NIS2 require traceable human accountability for incident classification. The EU AI Act classifies cybersecurity tooling in critical infrastructure as high-risk.

What the Field Had Actually Built

We ran a systematic review on LLM agents for cybersecurity with 4 core findings.

50.8%

of the corpus is penetration testing.

Incident response and forensics: 5 studies.

1

study on compliance and risk — also

the only banking-specific system.

0

of 59 studies are human-overseen.

52 are fully autonomous.

0

reach production-grade evaluation.

The field rests on lab assessment.

The field's default posture is the opposite of what a regulated deployment can accept.

Fully autonomous operation stands in structural tension with DORA's and NIS2's requirement for traceable human accountability.

Four Challenges That Follow

AUTHORITY

What may this agent reach?

An agent that can retrieve across the whole institution is a data-protection incident waiting for a prompt to trigger it.

ATTRIBUTION

On what basis did it conclude that?

“The model assessed this as critical” is not an audit trail, and will not survive a supervisory review.

OVERSIGHT

Who decides, and when?

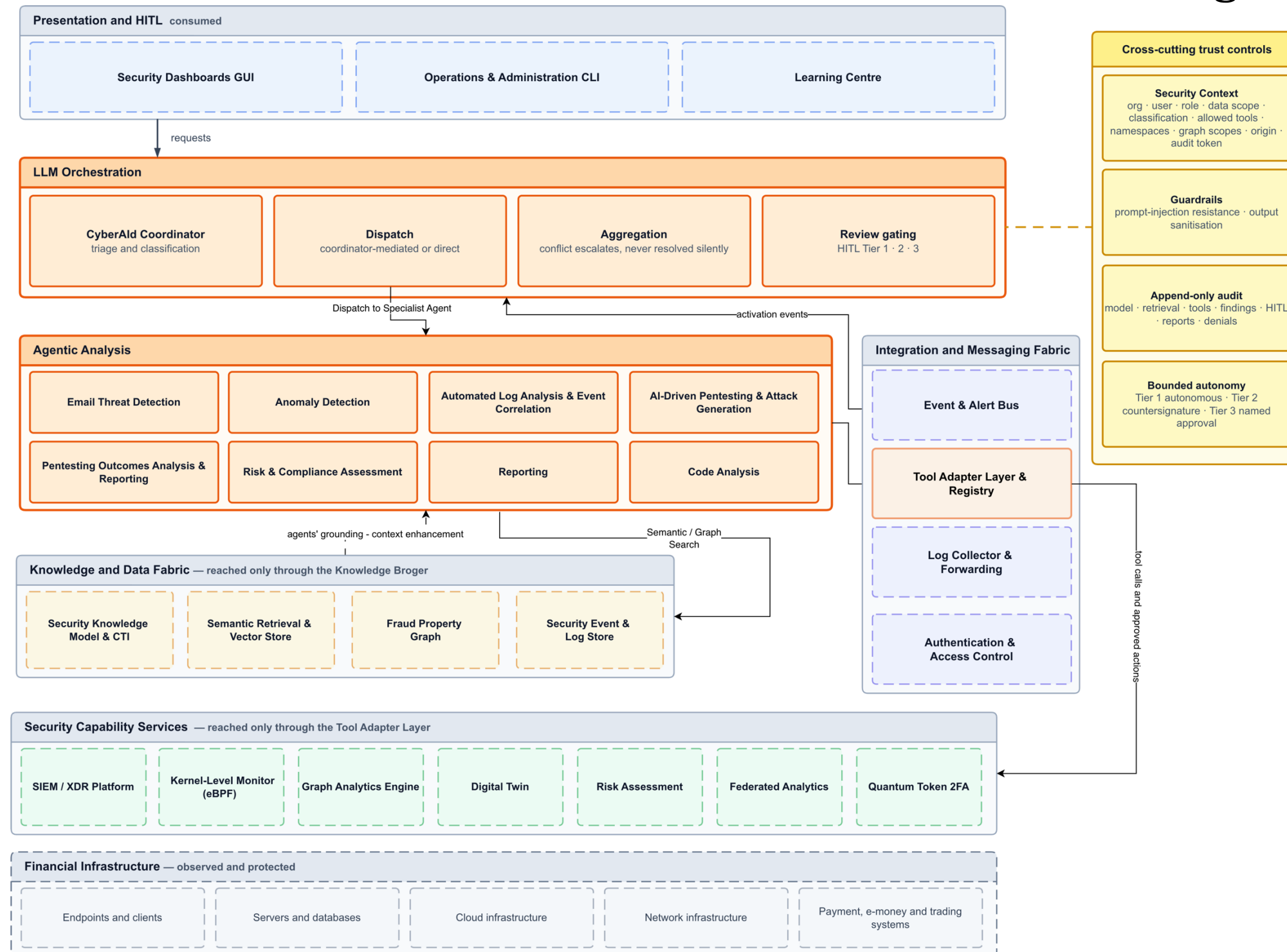
Approval for everything produces a queue nobody reads.

GROUNDING

Is the narrative true to the evidence?

A nice summary that outruns the evidence it cites is the failure mode of LLMs.

Where the Reasoning Layer Sits



Authority: One Security Context

Three properties govern how it behaves

Immutable for the workflow lifetime

narrowing can only remove — never add a namespace, a tool or a higher ceiling

Absence is a denial, not a default

no context, or an expired one, and the workflow does not execute

Enforcement is centralised

one component to review — agents do not implement their own checks

Security Context created at every entry point · carried for the whole workflow

schema_version

org_id

user_id

role

data_scope

classification

allowed_tools

rag_namespaces

graph_scopes

workflow_origin

audit_token

created_at / expires_at

extensions

one object, checked at eight boundaries

Agent invocation

role · allowed_tools

Agent denied

Document retrieval

rag_namespaces · classification

Retrieval denied

Graph query

graph_scopes

Graph query denied

Tool call

allowed_tools · argument schema
· operator path grants

Tool call denied

Memory read / write

org_id · data_scope

Memory access denied

Finding generation

classification · evidence-
reference policy

Finding denied

Report generation

finding-set scope · regulatory
template scope

Report denied

UI visibility

role · data_scope

Output redacted

Permitted

the operation proceeds and is recorded as a normal audit event

Append-only audit log

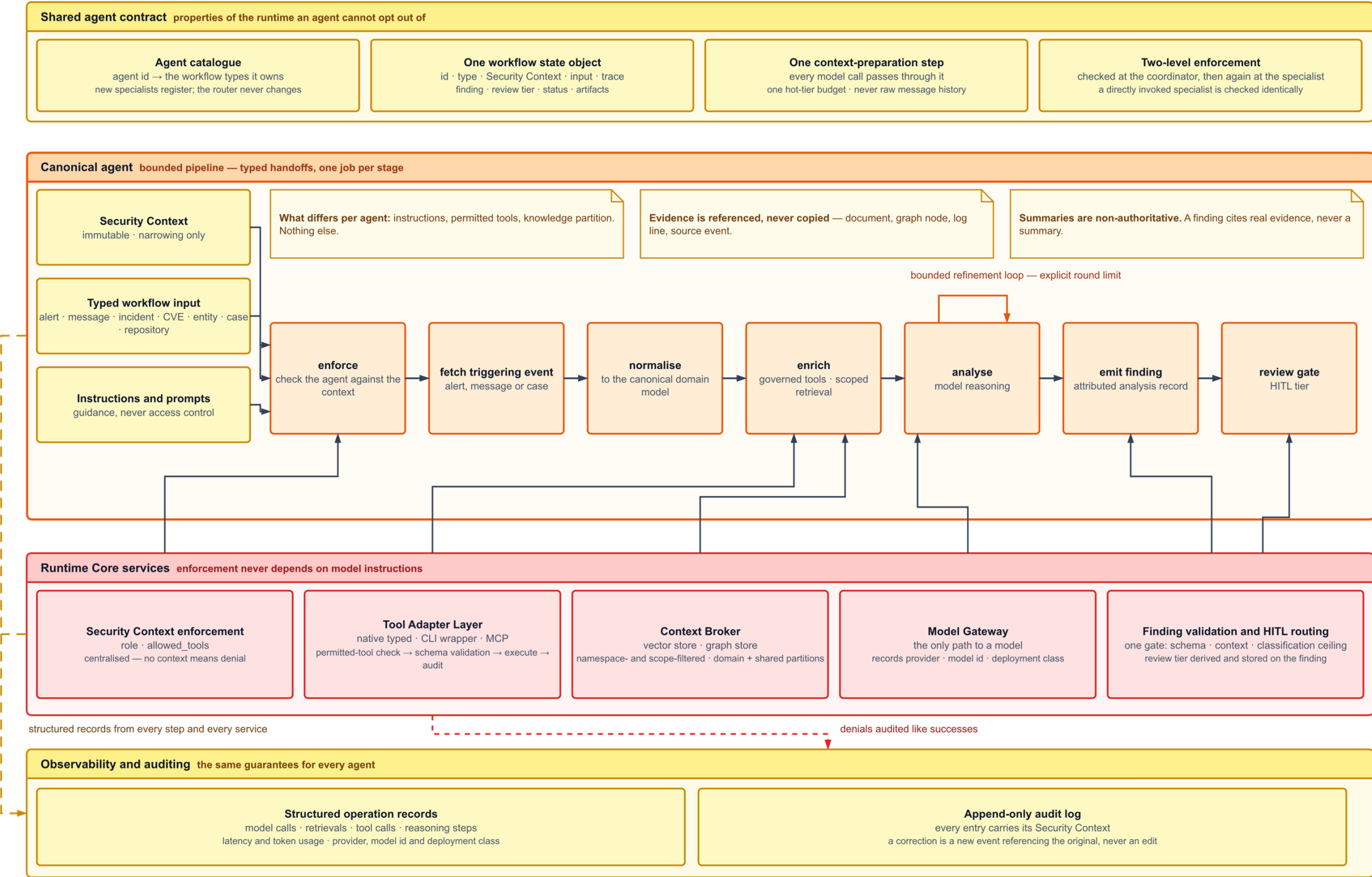
every denial names the invariant it violated
repeated out-of-scope attempts form a visible pattern, not an absence of results

Two extension paths

the extensions field adds dimensions without a schema change
policy plugins run after the core checks — tighten, never relax



Attribution: One Canonical Agent



Oversight: Proportionate, Not Uniform

The depth of review is tied to the reversibility of the action — because reversibility bounds the cost of being wrong.

Tier 3 High severity, low reversibility, customer-affecting, regulatory-binding
Named approval with recorded justification

Tier 2 Confidence-gated, medium impact, reversible containment, drafted notification
Analyst countersignature

Tier 1 Reversible enrichment, triage, summarisation, indicator lookup
None — proceeds autonomously

Tier 2 and 3 pause before the action

Grounding: The Finding Is the Contract

Every agent emits one validated record type. Nothing reaches a human by another path.

Identity

Authority

Assessment

Evidence

Action

Regulatory

Extension

Evidence is referenced, never copied.

A locatable source — an event, a file and line range, a document section, a graph node. Copying regulated content into findings would spread sensitive data into a new store with its own retention problem.

Claims must re-ground.

Assertions must trace to the references the finding carries. A finding whose narrative outruns its references is the failure this rule exists to catch.

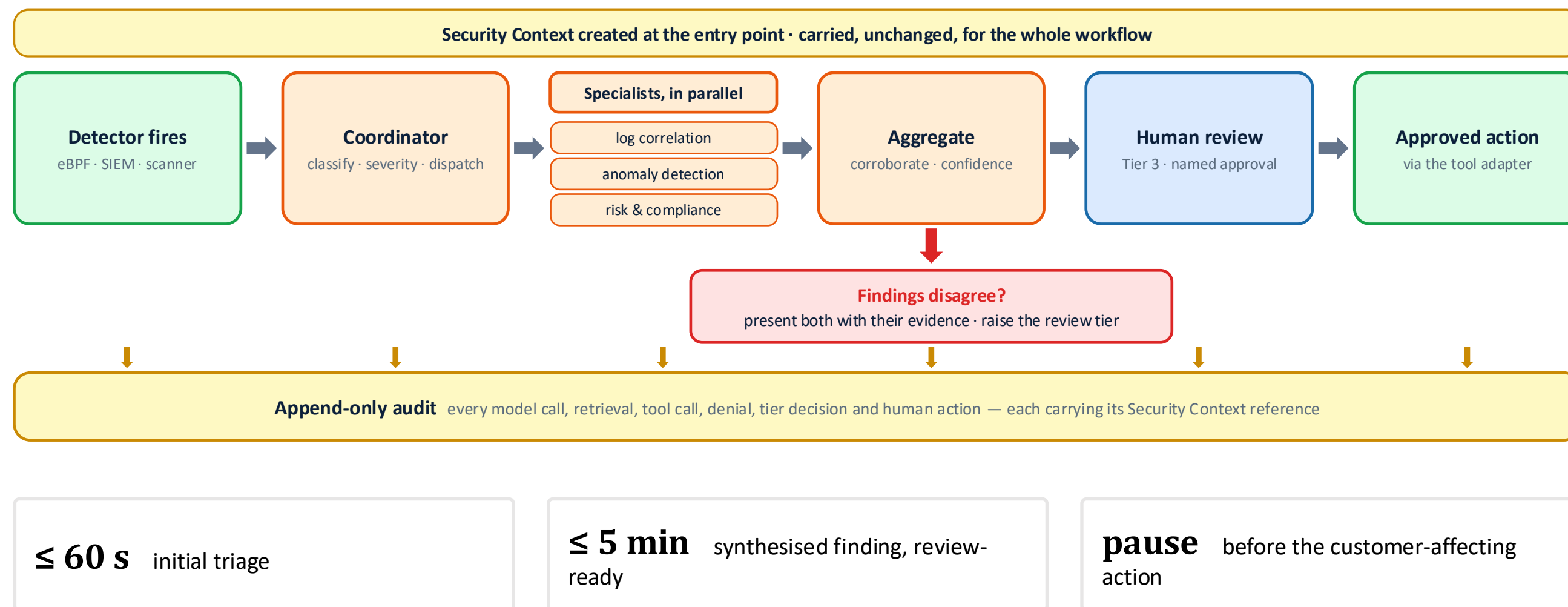
Validation is a single gate.

One service checks the schema, re-resolves the context, applies the classification ceiling, writes and audits. A malformed output is a visible rejection.

A reviewer reading a finding six months later has the same material the analyst had at the time — without re-running the workflow.

Regulatory mappings travel on the finding: DORA · NIS2 · GDPR · AML · EU AI Act · ISO 27001 · NIST CSF

Detection → reasoning → human decision → governed action. Every step reconstructable.



When specialists disagree the conflict escalates and both readings are presented with their evidence and the review tier is raised.



The Same Architecture, Sliced Four Ways

Each pilot selects the capabilities it needs

Client impersonation

Private banking & trading

Email threat detection + privacy-preserving federated analytics

> 90% accuracy on simulated impersonation

Anti-money laundering

Payment service providers

Fraud property graph + graph analytics + correlated log analysis

> 60% less investigation time

Anomaly detection

Retail banking

Multi-channel telemetry + orchestration + risk engine

> 50% reduction in time to respond

Cyber resilience

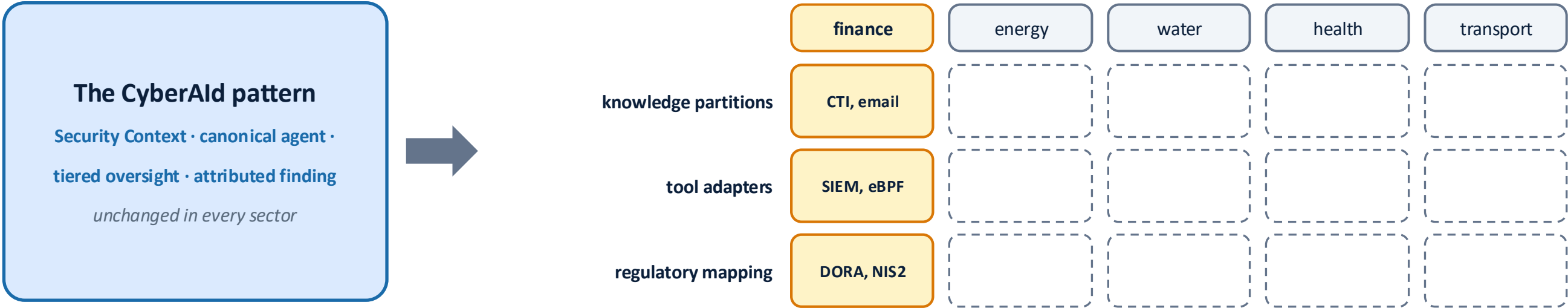
High-frequency trading

eBPF integrity monitoring + code analysis + feed integrity

< 5 ms added trading latency

What Transfers Beyond Finance

Three of the four answers are sector-agnostic. The fourth is a lookup table.



the pattern is the same — the three slots are what your sector fills in

Transfers unchanged

Security Context · canonical agent contract · tiered oversight · attributed finding

Sector-specific — and only these three

knowledge partitions · tool adapters · the regulatory mapping on the finding



Thank you!

Connect With CyberAid:

-  cyberaidproject.eu
-  CyberAid
-  CyberAidProject
-  Cyberaid
-  CyberAid

Research works

LLM-Based Agents for Cybersecurity: A Systematic Review
J. Cybersecurity and Privacy, 2026

An Organization-Scoped LLM Agent Runtime Architecture
IEEE DCOSS-IoT 2026

CyberAid: AI-Driven Cybersecurity for Financial Service
Providers IEEE DCOSS-IoT 2026



The project funded under Grant Agreement No 101249596 is supported by the European Cybersecurity Competence Centre

George Fatouros
gfatouros@innov-acts.com