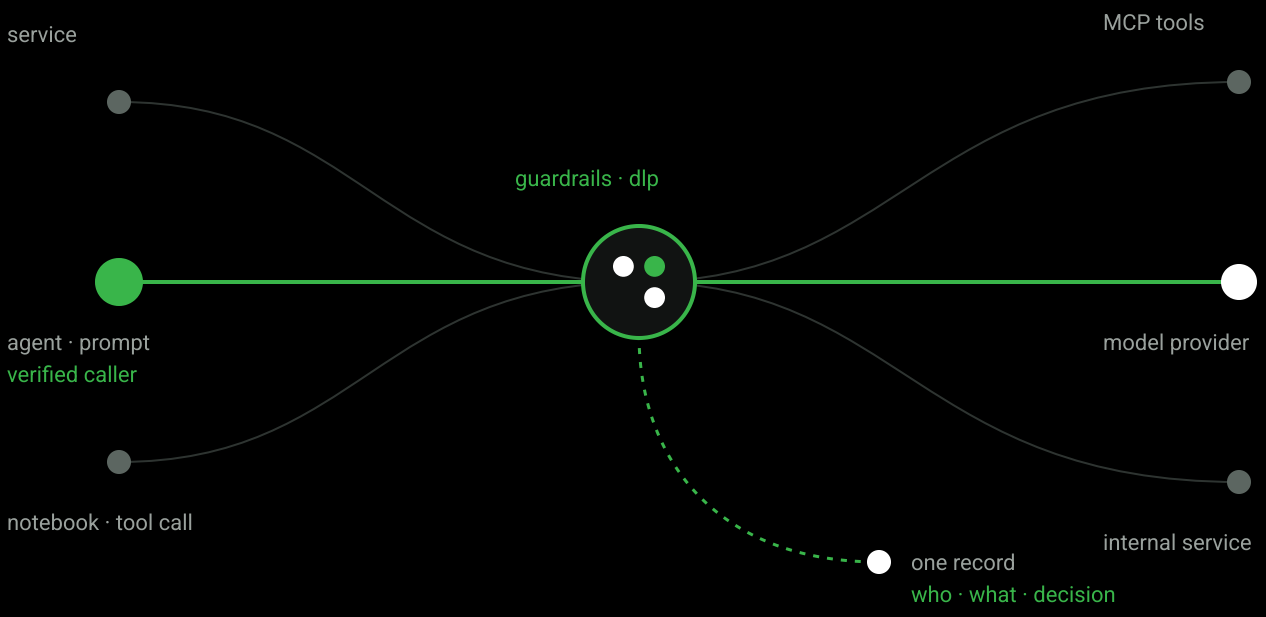
Greymatter.io

WHITE PAPER
GREYMATTER 2.5.0 · NOVEMBER 2026



Guardrails and DLP: inspect every prompt, tool call and answer where it runs

How Greymatter inspects AI traffic beside each workload, maps it to the OWASP Top 10 for LLM Applications, and names the sender.

© 2026 Greymatter.io, Inc.

Contents

		SECTION	PAGE
THE CASE	1	Executive summary	1
	2	AI traffic moved inside	3
	3	One pass, two products	4
INSPECTION	4	Guardrails: measure, then enforce	5
	5	DLP: evidence by caller	7
	6	Detection built to resist evasion	9
	7	Streams, sockets and images	10
	8	MCP and tool calls	11
COVERAGE	9	Mapping to the OWASP Top 10	12
	10	Inspection on every workload	13
	11	The record: one event per call	14
PROOF	12	Evidence for AI governance	16
	13	Limits and failure behavior	18
	14	How this differs	19
	15	Operational and mission benefits	20
	16	Ask Greymatter	21

How to read the status marks

Every capability in this paper carries one mark. Nothing marked Fielded is aspirational.

FIELDED

Ships in Greymatter 2.5.0 today.

PREVIEW

Ships in 2.5.0 as a technology preview.

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

1

Executive summary

Greymatter inspects every prompt, tool call and model answer inside the data plane that already carries the traffic. It needs no new gateway, add-on or product. It adds no hop, and agents do not change.

IN BRIEF

Two products share one inspection pass. Guardrails detects attacks on AI systems and can block them. Data loss prevention (DLP) finds sensitive data and records it.

WHO THIS IS FOR

Security architects, chief information security officers, program offices and assessors who must show that AI traffic is inspected, attributed and on the record.

AI moved inside the enterprise. Agents call models, models call tools, and tools call services. Most of that traffic never crosses a perimeter. A gateway at the edge inspects little of it. A scanning service off to the side inspects only what someone sends it.

Greymatter puts the inspection where the traffic already is. Its data plane runs beside every workload in the mesh. Guardrails and DLP run inside that data plane, on the way in, on the way out, or both. They run after Greymatter has verified who is calling. Every finding carries that identity.

Guardrails maps its detectors to the OWASP Top 10 for LLM Applications. It starts in detect mode and reports which requests it would block. When the results match what the team expects, it turns on enforcement one module at a time. DLP finds personal data, health data, credentials, logistics data and classification markings. It never breaks a business flow. It builds the evidence.

Greymatter writes every decision to one record that names who sent the traffic and what the policy did. That record is the runtime evidence that AI governance frameworks require. Two examples are the OECD AI Principles and ISO/IEC 42001.

THE CENTRAL POINT

Inspection that lives in the data plane covers the traffic a gateway never reaches, and it names who sent it.

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

Six properties a program office can test

Inside the data plane

FIELDED

Guardrails and DLP run in the data plane beside each workload. Traffic takes no detour to a separate inspection service.

Identity on every finding

FIELDED

Inspection runs after Greymatter verifies the caller. Each finding names the workload and, where known, the person.

Measure, then enforce

FIELDED

Detect mode reports what enforce mode would block. Each Guardrails module turns on enforcement on its own.

All ten OWASP risks

FIELDED

Guardrails and DLP map their modules to every entry in the OWASP Top 10 for LLM Applications, 2025.

Prompts, tools, streams

FIELDED

Inspection covers HTTP bodies, Model Context Protocol (MCP) tool calls, WebSocket and Socket.IO messages, and streamed answers.

No raw secrets on file

FIELDED

The record fingerprints or withholds every credential. No setting can make it store one in the clear.

STATED PRECISELY, SO IT IS NOT OVER-READ

DLP detects and records. It does not block and it does not redact. Guardrails blocks only in enforce mode, and only for the modules a team sets to enforce.

Guardrails inspects text, including text carried inside images. Every limit has a defined result, and the record shows it. Section 13 states each one.

PART OF A SERIES

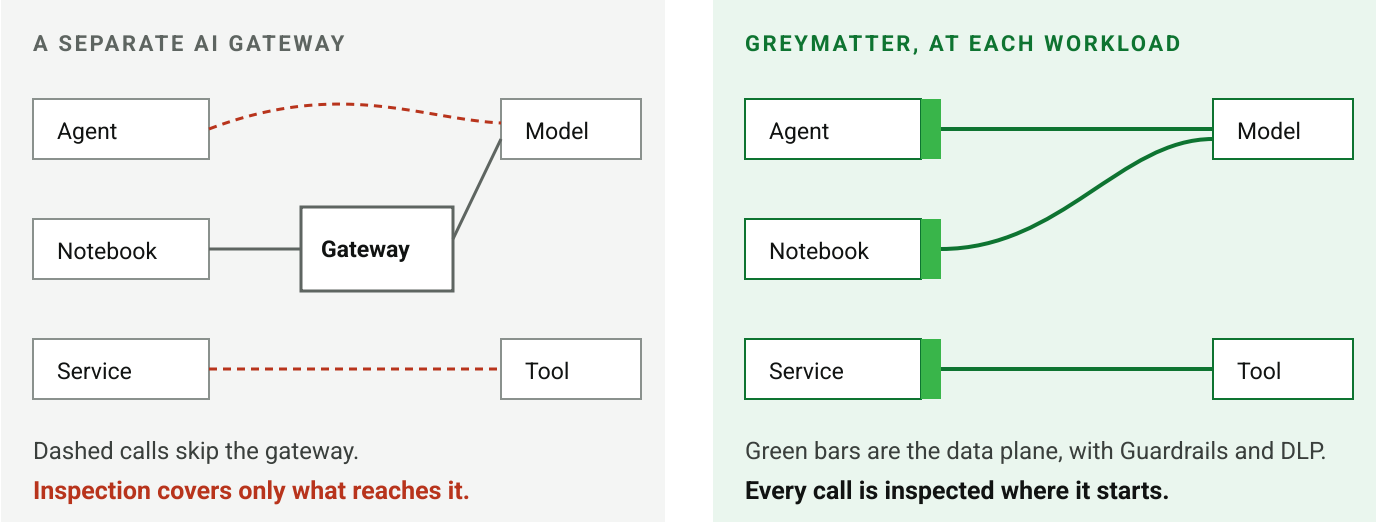
This paper goes deep on inspection. Section 12 maps it to AI governance frameworks. Companion papers cover verifiable identity, governed AI and autonomous onboarding.

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

2 AI traffic moved inside

Agents, notebooks and services call models and tools from inside the enterprise. Inspection has to be where those calls start.

FIGURE 1 · WHERE INSPECTION HAPPENS



THE SEAM

Every detour to a separate inspection point is a hop that someone must route traffic through, secure and keep running.

THE COST

A separate product also means a separate policy, a separate log, a separate team to operate it, and more open seams.

The common answers each leave a gap. A gateway at the edge inspects what crosses the edge. Internal calls between an agent and a tool never reach it. A scanning service inspects what an application sends to it, so every application must change, and every team must remember to call it. A fleet of separate AI gateways adds hops and seams, each with its own policy and log.

Each of those models also loses the caller. By the time traffic reaches a separate inspection point, it often carries a shared service credential. The finding says what was in the prompt. It cannot say which workload or person sent it.

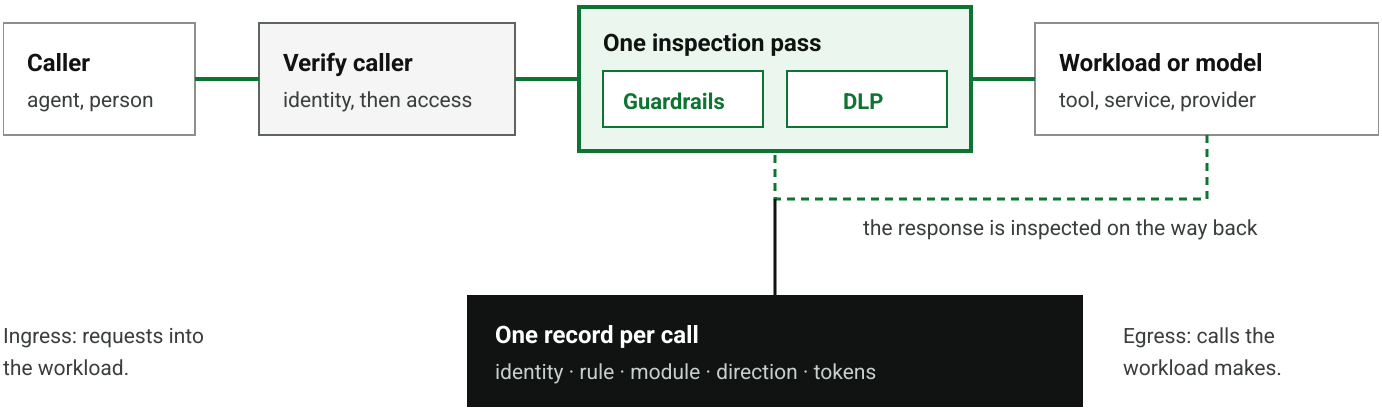
Greymatter starts from the other side. The data plane is already beside every workload in the mesh, with the caller's identity already verified. Guardrails and DLP run there.

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

3 One pass, two products

Guardrails and DLP share one read and one scan of each message. Each product keeps its own mode, its own limits and its own findings. FIELDED

FIGURE 2 · THE INSPECTION PASS IN THE DATA PLANE



	GUARDRAILS	DLP
Purpose	Find attacks on AI systems and stop them	Find sensitive data and record it
Modes	Detect, enforce or off	Detect or off
Can block	Yes, in enforce mode, per module	Never
Modules	Nine, mapped to OWASP LLM01 and LLM03 to LLM10, plus image inspection	Four: disclosure, protected health information (PHI), logistics and sensitive markings, mapped to LLM02
Default	Detect	Detect
Findings	Tagged Guardrails in the record	Tagged DLP in the record

PLACEMENT

Inspection runs after authentication. An unauthenticated request is refused before it costs a scan.

A team can turn on one product or both. When both are on, the data plane reads each body once and runs both sets of modules over it.

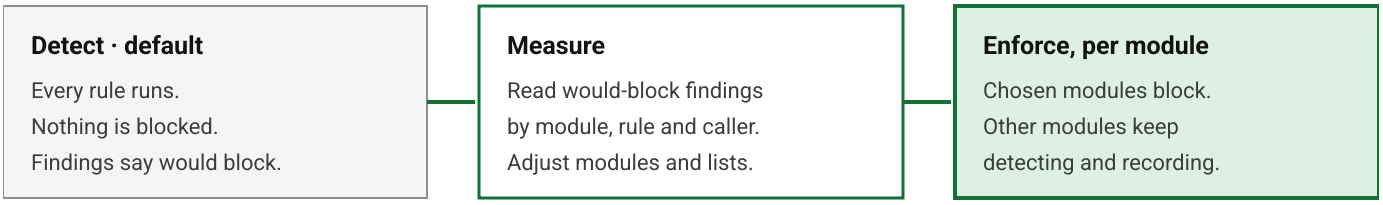
© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

4 Guardrails: measure, then enforce

Guardrails reports what it would block before it blocks anything. A team turns on enforcement one module at a time. FIELDED

FIGURE 3 · THE PATH TO ENFORCEMENT

FROM EVIDENCE TO ENFORCEMENT



DLP stays at detect. Its findings feed the record and never block.

DEFAULT

Every module starts in detect mode. No module blocks until a team sets it to enforce.

PER MODULE

Each module is off, detecting or enforcing. Prompt injection can enforce while misinformation only reports.

In **detect** mode, Guardrails runs every rule and changes nothing. It marks each finding that enforce mode would block. A team can rank callers, routes and rules by those findings before any traffic is at risk.

In **enforce** mode, the modules a team set to enforce can block. The other modules keep detecting. **Off** stops evaluation.

A team controls Guardrails in its filter definition, with no new release. It sets each module to off, detect or enforce. It can turn on lists of approved tools, retired tools and approved MCP methods. A call that breaks one of those lists is blocked when the agency module enforces. Lists of approved hosts and packages raise findings only.

Who may use a tool is set apart from inspection. The MCP Router grants tools to roles read from the caller's verified token. Mesh policy grants which workloads may call which services.

When a block takes effect

TRADE

Holding data adds the scan time to the request. Most chat traffic does not need it. Agents that act on tool calls often do.

By default, the data plane forwards data as it arrives and applies a block as soon as the scan finishes. A team can choose to hold data until the scan decides. Then a block takes effect before the service or model acts on the request, or before the client gets the answer.

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

WHY SCORES

A single weak signal is often innocent. Several weak signals in one request are not. The score catches that pattern.

Strong signals block. Weak signals add up.

High-precision rules can block on their own. Weaker rules add to a score instead, and the score blocks when enough weak signals add up in one call.

On a long-lived connection, such as a WebSocket, a team chooses whether Guardrails scores each conversation or the whole connection.

What a blocked caller receives

CALLER	REPLY
Web or API client	HTTP 403 that names the rule and the detector's message, never the matched text
MCP client	A JSON-RPC error with the rule number, so the client reads a protocol error
gRPC client	The matching gRPC status
Streamed answer	The stream resets, or ends with an error event, as the team chooses

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

5

DLP: evidence by caller

DLP finds sensitive data in prompts, tool calls and answers, and records who sent or received it. It never blocks a business flow. FIELDED

DLP has four modules. Each one is on or off. All four run in the same pass as Guardrails and add no second scan.

DLP is for the questions an assessor asks after the fact. Which workloads sent personal data to an outside model? Which person pulled patient identifiers into a notebook? Did any answer carry a classification marking? The record gives the answer by caller, by direction and by rule.

WHY DETECT ONLY

A false block on a valid patient record or shipping manifest stops a mission. DLP builds the record. Guardrails does the blocking.

BOTH DIRECTIONS

DLP reads requests and responses. A leak in a model's answer is a finding, the same as a leak in a prompt.

MODULE	WHAT IT FINDS
Disclosure	US Social Security numbers, payment card numbers, IBAN and US bank routing numbers, US tax IDs, Canadian Social Insurance numbers, passport and driver license numbers. European national IDs: UK National Insurance numbers, German, Spanish and Austrian social security numbers, French NIR, Italian codice fiscale and Danish CPR numbers. Dates of birth, email addresses, phone numbers and home addresses. Special categories under the EU General Data Protection Regulation (GDPR): race, religion, political opinion, union membership, sexual orientation and biometric data. Cloud access keys, code host and chat tokens, payment API keys, JWTs and private keys.
PHI	ICD-10, CPT and HCPCS codes, DEA numbers, NDC codes, LOINC codes and Medicare beneficiary IDs. Patient-linked diagnosis, medication and lab narrative. FHIR Patient resources. Clinical trial IDs, device identifiers and genomic variants.
Logistics	HAZMAT UN numbers, bills of lading, container numbers and customs codes. Fleet GPS and hours-of-service data. Requests to extract fleet telematics or infrastructure schematics, for military, first response and critical infrastructure programs.
Sensitive markings	Classification markings in the format set by the Controlled Access Program Coordination Office (CAPCO): banners, portion marks and control markings. Attempts to elicit classified information or declassification metadata.

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

Built to keep false positives out of the record

WHY IT MATTERS

A record full of false positives is a record nobody reads. Precision is what makes DLP evidence usable.

DLP reports a value only when it is valid for its kind, not when it merely looks like one. A number that fails its official check, such as the check digit on a card or a national ID, is not a finding.

As an example, DLP does not report a random short number as a card security code. A classification word in ordinary prose is not a marking. The result is a record an assessor can trust.

From a question to a leak

TWO FINDINGS, ONE STORY

The record links the request for the data and the answer that delivered it.

DLP links a request and its answer. When a request tells a model to list identities or dump patient records, DLP marks it as an extractive request. When the answer then carries personal data, health data or a secret, DLP raises an outcome finding that ties the two together.

DLP also compares the classes of data in the answer with the classes in the request. If an answer holds a kind of personal data that the request did not, the model or a tool introduced it. That is a leak, and the record says so.

SINGLE-USER WORKLOADS

A workload that runs for one person, such as a personal analysis server, makes outbound calls that carry no inbound request. A policy can name the annotation that holds the owner. Then every DLP and Guardrails finding on those calls names that person.

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

6

Detection built to resist evasion

Attackers change the spelling, the encoding and the wrapper. Guardrails reads past all three before it matches. FIELDED

READS WHAT IT INSPECTS

Guardrails reads tool calls, model files and links as what they are, not as loose text.

Attackers disguise what they send. They hide instructions in invisible characters, encode them, or put them inside a file or a tool call. Guardrails removes the disguise before it checks a rule, so the disguised attack and the plain one get the same result. The checks run inside the data plane, so no content leaves it.

WHAT ATTACKERS TRY	WHAT GUARDRAILS DOES
Encoded or compressed content	Unpacks it before the check
Invisible characters	Removes them, so hidden instructions show
Lookalike letters	Treats letters that look alike as the same letter
Random-looking text	Flags strings built to steer a model
Unsafe model files	Flags model files that would run code when loaded
Risky tool calls	Flags calls that are destructive, open-ended or far too large
Deceptive links	Flags links in answers that hide their real destination
Another tenant's data	Flags retrieval requests for it, checked against the caller's verified identity

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

7 Streams, sockets and images

AI traffic is rarely one request and one response. Guardrails and DLP follow it into streams, sockets and images. FIELDED

TRAFFIC	WHAT GREYMATTER INSPECTS
Requests and responses	Prompts, tool results and model answers, in the formats AI traffic uses. A team chooses which response types to inspect.
Streamed answers	Answers that a model writes word by word, held or checked as they flow
WebSocket and Socket.IO	Each message in chat apps and agents that keep a connection open
MCP	Each tool call, with its tool and its arguments. See section 8.
Images	Instructions and secrets hidden in images, and images that hide programs

HOW A STREAM ENDS

A team chooses whether a blocked stream resets, or ends with an error event that the client can read.

IMAGES ARE OPT-IN

A team turns image inspection on for Guardrails. In enforce mode it can block an image that hides a program.

Streamed answers force a trade between safety and speed. Greymatter lets each team make that trade for each workload, in its Guardrails settings. Buffer mode holds the whole answer until the check ends, so nothing unchecked reaches the user. Passthrough mode streams the answer as the model writes it and checks it along the way. When Guardrails blocks a stream in enforce mode, the stream stops at once.

Images are a growing path for prompt injection and data loss. An attacker can hide instructions in an image that a user uploads. An image can also carry a secret in its embedded text. With image inspection on, Guardrails checks the text an image carries with the same prompt-injection and disclosure rules as any other text.

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

8

MCP and tool calls

An agent that can call tools can act. Guardrails reads each tool call as a protocol message, not as plain text. FIELDED

PROTOCOL AWARE

Guardrails decodes each JSON-RPC request, including batches. Rules read the method, the tool and every argument.

KEY SMUGGLING

Two parsers can read a repeated JSON key differently. Guardrails flags repeated and ambiguous keys, so a hidden argument cannot slip through.

With MCP inspection on, Guardrails reads which method a client called and which tool it requested. It reads the arguments as values, so a prompt injection hidden in one argument is still found. It also checks the tool lists that a client offers to a model.

When Guardrails blocks an MCP request, the client receives a JSON-RPC error that carries the rule number. The agent receives a protocol error it can handle, not a web page.

CONTROL	WHAT IT DOES
Method allowlist	Flags a JSON-RPC method that is not on the list, and blocks it in enforce mode
Tool allowlist	Flags a tool that is not approved, and blocks it in enforce mode
Tool denylist	Flags a tool the team has retired, and blocks it in enforce mode
Destructive tools	Flags a call to a tool that deletes, writes or runs commands
Open schemas	Flags tool schemas that accept any extra property
Oversized arguments	Flags arguments far larger than a tool call needs

DENIED BY DEFAULT

A caller without a grant for a tool gets a not-permitted error. Client credentials never pass to the tool server.

The MCP Router PREVIEW

The MCP Router gives agents one endpoint for many MCP servers. It grants tools per caller and limits calls per caller and per tool. A router the Greymatter CLI creates comes with Guardrails and DLP on its listener, in detect mode, with MCP inspection on. Every tool call through it goes on the record.

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

9 Mapping to the OWASP Top 10

Each entry in the OWASP Top 10 for LLM Applications, 2025, the module that covers it, and what that module looks for. FIELDED

ENTRY	RISK	MODULE	WHAT IT LOOKS FOR
LLM01	Prompt injection	Prompt injection	Direct and hidden instructions, smuggled characters, adversarial suffixes, injected tool calls
LLM02	Sensitive information disclosure	DLP: disclosure, PHI, logistics, sensitive markings	Personal, health, logistics and classified data, secrets, extractive requests and leaks in answers
LLM03	Supply chain	Supply chain	Unsafe model files, remote code flags, cloud metadata probes and credentials in requests
LLM04	Data and model poisoning	Poisoning	Planted few-shot examples, fake retrieval context, hidden text and trigger phrases
LLM05	Improper output handling	Output handling	Script and HTML injection, SQL, path traversal, suspicious links, unapproved packages
LLM06	Excessive agency	Agency	Unapproved and destructive tools, unlisted MCP methods, open schemas, oversized arguments
LLM07	System prompt leakage	Prompt leakage	Probes for the system prompt, and prompt fragments, keys and tokens in answers
LLM08	Vector and embedding weaknesses	Embedding	Raw vector returns and cross-tenant retrieval, checked against verified identity
LLM09	Misinformation	Misinformation	Unsafe code, weak cryptography advice, invented package names, false claims of expertise
LLM10	Unbounded consumption	Consumption	Oversized messages, token and sample counts, deep nesting and repetition

BLOCKING

Every Guardrails module can block in enforce mode. The DLP modules for LLM02 record and never block.

Coverage means a module looks for that risk in traffic. It does not mean every form of the risk is caught. No traffic filter can settle whether an answer is true. For misinformation, Guardrails flags the signals it can measure.

BEYOND THE FILTER

Token budgets and rate limits also address LLM10. The Governed AI paper covers them.

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

10

Inspection on every workload

A team does not build inspection into each service. It writes one policy entry, and Greymatter applies it to every workload the entry matches. FIELDED

COVERS NEW WORKLOADS

The entry matches by label, so a workload created tomorrow gets the same inspection as one running today.

OWNER ON OUTBOUND CALLS

An entry can name the annotation that holds a workload's owner, so findings on its outbound calls name that person.

Inspection is part of the Tenant Onboarding Policy that governs a namespace. A **filters** entry selects workloads by label and names the filter that holds the Guardrails and DLP settings. The entry also says which direction to inspect.

Ingress inspects the requests that reach the workload through the mesh. **Egress** inspects the calls the workload makes, and the answers it gets back, such as calls to a model provider. An entry can choose both.

The inspection runs in the workload's own data plane. It runs after Greymatter resolves the caller, on every mutual TLS path into the workload. Nothing in the application changes.

WHERE INSPECTION ATTACHES	HOW A TEAM TURNS IT ON
Every workload a label matches	A filters entry in the Tenant Onboarding Policy, inbound, outbound or both
A service defined in configuration	A DLP or Guardrails definition on its listener, in Greymatter Specification Language (GSL) or a custom resource
A route out to a provider	The same definitions on the egress route that reaches the provider
An MCP Router	Created with Guardrails and DLP in detect mode by the Greymatter CLI

SAFE CHANGES

The data plane checks every setting strictly. It refuses a setting it cannot run and keeps its last good configuration.

The Console reports each product per listener, so an operator can see at a glance where inspection is off.

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

11

The record: one event per call

Every finding from Guardrails and DLP goes to the same record as the rest of the platform's evidence, with the caller attached. FIELDED

ONE PLACE

The platform's log store collects every event. A service defined in GSL can also send events to an OpenSearch index or a rolling log file.

READABLE

Each event carries a short plain-language summary of what was found, for the analyst who is not a rule author.

An event describes one call: a request and its response. It names the caller and what the caller sent. It lists every finding, what the policy did, and what the call cost in tokens. An assessor can filter by person, workload, module, rule or direction, without joining logs from separate products.

THE EVENT RECORDS	DETAIL
Who	The verified workload identity and, where known, the person or the workload owner
Where	Namespace, service and listener, client and host addresses, method and path
What was decided	Mode, the action taken, and the rule that blocked, if any
Each finding	Rule, module, product, direction, severity, whether it would block, and a short match
Scores	The anomaly scores at the time of the decision
Streams	Transport, and the conversation a socket message belonged to
Model use	Input, output and total tokens, the model, and whether the provider or Greymatter counted them
Bodies	The start of each body, on by default and limited in length, so a team can turn it off

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

No raw credential ever reaches the record

ALWAYS ON

This protection has no switch and no setting. No configuration can make Greymatter write a raw credential.

An inspection record that stores the secrets it found becomes a leak of its own. Greymatter treats every credential in an event, in a header, a path or a body, by one rule. Keep what says who, what, when and how. Remove what could be replayed.

REPLAY EVIDENCE

The same fingerprint on two requests means the same token was used twice. That alone can show a replay.

KIND OF VALUE	STORED AS
User names, key IDs, client IDs, token subjects	Kept, because they identify and cannot be replayed
Bearer tokens, session IDs, API keys without a public ID	A keyed fingerprint, so records link without the token
Passwords, one-time codes, client secrets, private keys	Withheld entirely, with no fingerprint

In the Console

GAPS IN VIEW

An engine that is off on a listener stays in view, marked off, so a gap shows at a glance.

Each service's view shows Guardrails, DLP and web application firewall findings for a listener, side by side. It shows a timeline, the rules that fired, and a summary of what stands out.

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

12

Evidence for AI governance

Every AI governance framework requires controls that run and records that prove they ran. Greymatter supplies both, at runtime, from the data plane. FIELDED

EXTEND, DO NOT BUILD

A team turns on Guardrails and DLP in the mesh it already runs. It builds no logging pipeline and buys no second gateway.

International AI frameworks share a pattern. Examples are the OECD AI Principles, ISO/IEC 42001, the NIST AI Risk Management Framework (AI RMF) and the EU AI Act. Each requires an organization to find where AI runs, set controls, watch the system, and keep a trail. Written policy is the simple part. Greymatter proves the policy ran on real traffic.

Defense programs add NATO's Principles of Responsible Use of AI in Defence. National principles, such as those of the United States, come last in the table.

FRAMEWORK	WHAT IT REQUIRES	WHAT GREYMATTER SUPPLIES
OECD AI Principles	Transparency and explainability	Every AI call on record, with caller, rule, model and tokens
	Robustness, security and safety	Guardrails blocks attacks in enforce mode. Coverage gaps show what was missed
	Accountability	Each finding names the workload and, where known, the accountable person
NATO Principles of Responsible Use of AI in Defence	Explainability and traceability	The record, written by the data plane, for every call
	Governability	Enforce mode blocks misuse. Policy turns a module, egress or tool off
ISO/IEC 42001	Operation and monitoring	Runtime controls and the records that show they ran
NIST AI RMF	Map, Measure and Manage	Which workloads call which models and tools, what they sent, and the controls
EU AI Act	Record-keeping, Article 12	Automatic logs of each use, kept by the platform rather than the application
United States Intelligence Community AI ethics	Transparent and accountable. Secure and resilient	The same record and controls, with the verified caller on each finding
United States DoD AI Ethical Principles	Traceable and governable	The record, and enforcement a team can turn on or off per module
United States CDAO RAI Toolkit, SHIELD	Log for traceability. Detect via continuous monitoring	The record, and Guardrails and DLP findings on live traffic

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

RUNTIME, NOT PAPERWORK

Greymatter covers the part of governance that happens in live traffic. It feeds the systems that hold the rest.

NOT A CERTIFICATION

Using Greymatter does not make a system compliant with any framework. It supplies controls and evidence an assessor can test.

Where the line sits

A governance program has more parts than any one product covers. Greymatter does not keep a model registry, test a model for bias, write model cards or route approvals. It does not decide whether a use case is acceptable. Those belong to the governance and risk systems an organization already runs.

What those systems lack is evidence from the traffic itself. Greymatter supplies it. The record goes to the platform's log store, and services defined in GSL can also send it to an OpenSearch index. A risk team reads which agents called which models, what they sent, what Guardrails and DLP found, and what the policy did.

GOVERNANCE NEED	WHO MEETS IT
Inventory of AI use on the network	Greymatter, from the traffic it carries
Controls on prompts, tools and answers	Greymatter Guardrails, in enforce mode
Evidence of sensitive data in AI traffic	Greymatter DLP
Cost and usage per person	Greymatter token budgets and the record
Model registry, bias testing, model cards	The organization's model governance tools
Use case approval and risk acceptance	The organization's governance and risk process

FAIRNESS AND BIAS, STATED PRECISELY

The OECD fairness principle and the NATO bias mitigation principle require steps to reduce unwanted bias. Traffic inspection cannot measure bias in a model. Greymatter flags some signals of a bad answer, such as unsafe code or invented packages, but bias testing belongs to model evaluation.

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

13

Limits and failure behavior

Inspection must never become the outage. Each limit in the following table has a defined result, and each result is on the record. FIELDED

WHEN THIS HAPPENS	GREYMATTER DOES THIS
A scan runs past its deadline	Lets the traffic through and records a coverage gap. A team sets the deadline for each product.
A body is larger than the scan size	Scans up to the size the team sets and records that the rest was not scanned. In enforce mode, a team can block such bodies instead.
A setting is invalid	Refuses it and keeps the last good configuration.
DLP finds sensitive data	Records it. DLP never blocks.

WHY A SLOW SCAN LETS TRAFFIC THROUGH

A scan that hangs would hold every request behind it. A recorded gap is visible and fixable. A stalled data plane is an outage.

WHERE TO BE STRICT

Teams that must stop every unscanned body can set oversize to block and hold data until the scan decides, on the workloads that need it.

The coverage-gap records matter as much as the findings. They show an assessor exactly which calls were not fully inspected, and why. A team can watch them and raise a limit for the services that need it.

Guardrails reads text. It does not read pixels in images, audio or video. It does not decide whether a model's answer is true. It flags the signals of a false answer that it can measure.

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

14

How this differs

Zero trust means no implicit trust on any path. Inspection that lives on one path, or beside the path, leaves the others open.

ONE PRODUCT

Guardrails and DLP ship in the same data plane, policy, record and Console as the mesh. There is nothing extra to integrate or run.

WHAT A TEAM KEEPS

Identity on every finding, inspection on internal calls, and one record that does not depend on any one team's memory.

Three models are common. A team routes AI traffic through a separate AI gateway. A team calls a scanning service from each application. Or a team buys a separate gateway per model provider. Each model adds a hop, a seam and a policy. Each one covers only the traffic that someone sends to it.

Greymatter adds no hop. Inspection runs where the mesh already decrypts traffic and verifies the caller. The table compares the three approaches as a class. Products within a class vary.

	SEPARATE AI GATEWAY	SCANNING SERVICE	GREYMATTER
Where inspection runs	A separate hop that traffic must be routed through	An API each application must call	In the data plane beside each workload
Calls between agents and tools inside	Seen only if routed to the gateway	Seen only if the application sends them	Inspected where they start or arrive
Who sent it	Often a shared service credential	Whatever the application reports	The verified workload and, where known, the person
Application changes	New endpoints and routing	Code in every application	None
New seams	A new hop to secure and run	A new service to secure and run	None added
Policy and record	Separate from the mesh	Separate from the mesh	The same policy and record as the mesh
Streams, sockets and MCP	Varies	Varies	Inspected in their own formats

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

15

Operational and mission benefits

Each outcome, what produces it, and its status.

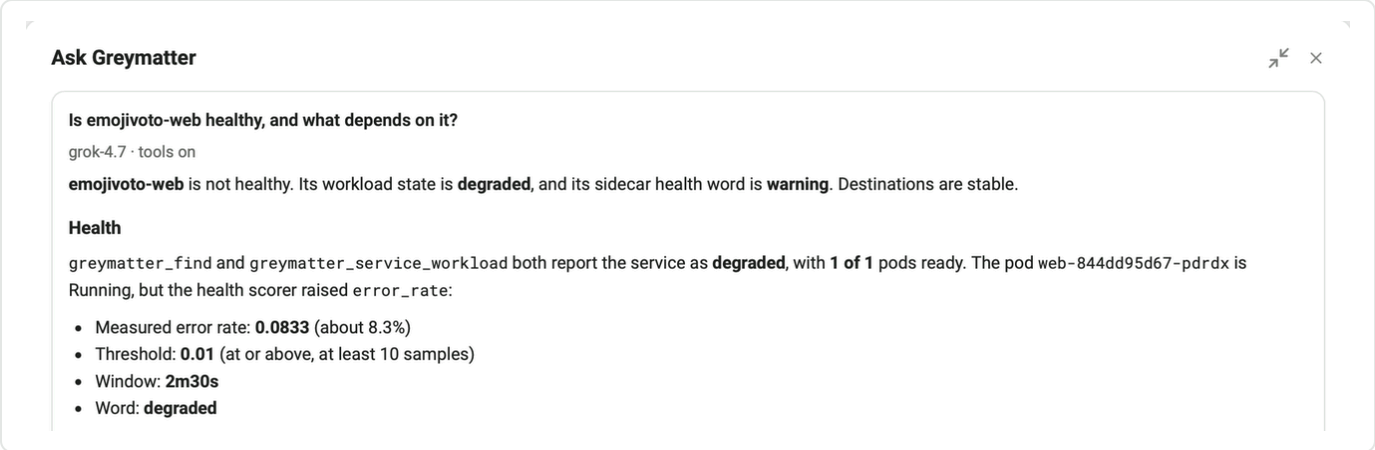
OUTCOME	WHAT PRODUCES IT	STATUS
Coverage without application changes	Inspection in the data plane beside every workload, by label	FIELDED
Findings tied to people	Inspection after identity, with the person or owner on each finding	FIELDED
Enforcement without surprises	Detect mode that reports what enforce mode would block	FIELDED
Precision a team can tune	Per-module enforcement, score stacking and allowlists	FIELDED
Evidence for LLM02	DLP for personal, health, logistics and classified data, never blocking	FIELDED
Coverage of the OWASP Top 10	Modules mapped to every 2025 entry	FIELDED
Agents under control	MCP inspection, tool and method lists, and protocol-correct blocks	FIELDED
Streams that stay fast	Passthrough scanning with bounded exposure	FIELDED
Hidden content found	Disguised content unpacked, and text read from images	FIELDED
A record that is safe to keep	Credentials fingerprinted or withheld, always	FIELDED
Gaps that are visible	Coverage-gap records for every call not fully scanned	FIELDED
One product to run	Same data plane, policy, record and Console as the mesh	FIELDED
Evidence for AI governance	Runtime records for the OECD, NATO, ISO/IEC 42001, NIST and EU AI Act frameworks	FIELDED
One endpoint for tools	MCP Router with grants, call limits and inspection	PREVIEW

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

16 Ask Greymatter

Ask Greymatter answers questions about what Guardrails and DLP found, where and for whom. FIELDED

FIGURE 4 · ASK GREYMATTER IN THE CONSOLE



READS AS YOU

Each tool reads the platform as the person who asked, and only reads. An answer holds only what that role allows.

GOVERNED LIKE ANY AI USE

Guardrails and DLP check each model call. The audit trail names the person and the agent.

Ask Greymatter is on every page of the Console. A person types a question in plain language. A model the organization chooses answers it from the platform's own data, through read-only tools.

For Guardrails and DLP, it reads where each engine runs, the findings on each service and the identities behind them. These tools need a security viewer role.

Tools come in groups: Health, Security, Traffic, Changes and Docs. Health and Docs are on by default. A person turns on the others in the Tools menu.

Questions a team can ask

“Which services had Guardrails or DLP findings today?”
`Security` · where each engine found something

“Show me the last 10 DLP findings on the orders service.”
`Security` · the findings, newest first

“Which paths on the orders service did Guardrails flag?”
`Security` · one service's top rules and paths

“Which identity did Guardrails flag most today, and what else did it do?”
`Security` · findings, then the request records

“How many DLP findings were there this week, against the week before?”
`Security` · totals with the previous period's

“How do I run Guardrails on both inbound and outbound traffic?”
`Docs` · the Greymatter documentation

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

Key takeaways

1

Inside the data plane

Guardrails and DLP run beside every workload, in the data plane that already carries its traffic.

2

Identity first

Inspection runs after Greymatter verifies the caller, so every finding names who sent it.

3

Measure, then enforce

Detect mode shows what enforce mode would block. Each module turns on enforcement on its own.

4

DLP builds evidence

DLP finds personal, health, logistics and classified data in both directions, and never blocks.

5

Tools and streams

Greymatter reads MCP calls, socket messages and streamed answers in their own formats, not as raw bytes.

6

A record safe to keep

One event per call, with identity and tokens, and no raw credential, ever.

IN ONE SENTENCE

Greymatter inspects every prompt, tool call and answer where it happens, names who sent it, and keeps a record an assessor can trust.

© 2026 Greymatter.io · Greymatter.io is a trademark of Greymatter.io, Inc.

Greymatter.io

Inspect where it runs. Name who sent it. Keep the record.

LEARN MORE

greymatter.io

Greymatter.io is a trademark of Greymatter.io, Inc.

© 2026 Greymatter.io, Inc.