

The Unguarded Prompt: Securing LLM Applications in Production

An Igris Security Whitepaper



IGRIS SECURITY
WHITEPAPER

The Unguarded Prompt

Securing LLM Applications in Production

Zero trust for AI.

Every prompt and every token: governed, audited, controlled.



Every prompt your users type is a data transfer to a third party. Can you say exactly what left your network today?

This whitepaper is for CISOs, CTOs, and the security and platform leaders who work with them in fintech, healthtech, legal tech, and enterprise SaaS. It explains where LLM risk really comes from, sets out eight controls that address it, and maps each control to the frameworks your auditors and customers already use.

1. Executive summary

Every LLM feature your company ships sends data to a company you do not control. Most security teams cannot say what left, who sent it, or whether it should have.

The answer is not to slow AI down. It is one checkpoint between your applications and every model, enforcing eight controls and producing audit evidence by default. This paper explains where the risk comes from, sets out those controls, maps them to frameworks such as NIST AI RMF, ISO/IEC 42001, the EU AI Act, GDPR, and India's DPDP Act, and gives you a quick way to score where you stand. Igris appears only in the final chapter.

2. The new data exit

LLM traffic has become one of the largest flows of sensitive data leaving the modern company, and usually the least watched. To a firewall it looks like ordinary HTTPS. Inside, it carries customer messages, retrieved documents, source code, and logs, and model responses can carry sensitive data straight back to users.

FIGURE 01

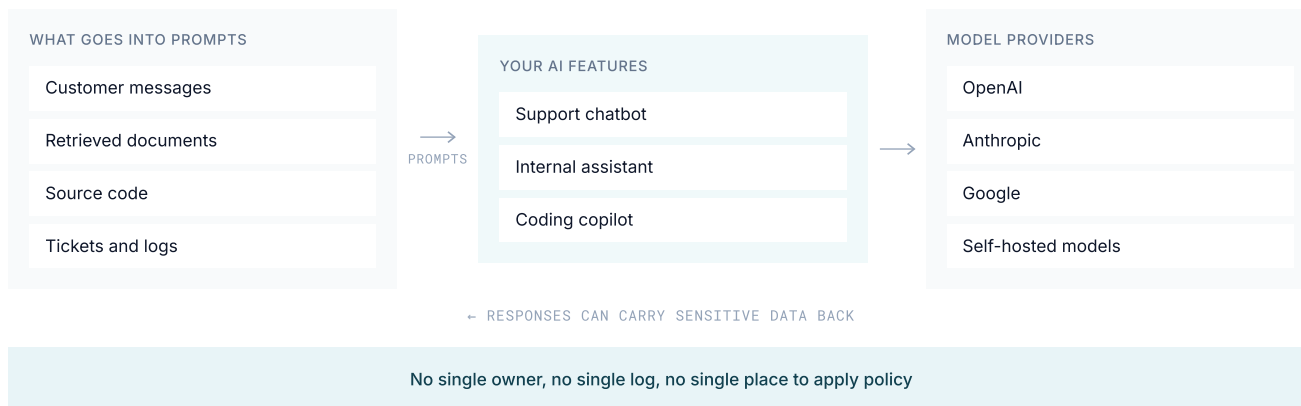
LLM traffic leaves through many paths, with no single log or policy

Figure 1 · How LLM traffic leaves a typical company

Sensitive text flows from many sources, through many AI features, to many providers, and responses flow back the same way. No single place sees all of it.

Three ways it happens every day

- **The helpful customer** types a full card number into a support chatbot, and the whole conversation lands in a provider's logs.
- **The fast-moving engineer** tests a new feature with real patient notes because the synthetic data looked too clean.
- **The invisible team** builds an assistant on a new provider with a personal API key, with no contract review and no log.

None of them are acting in bad faith. They are doing their jobs with the fastest tool available, which is exactly why the problem grows quietly.

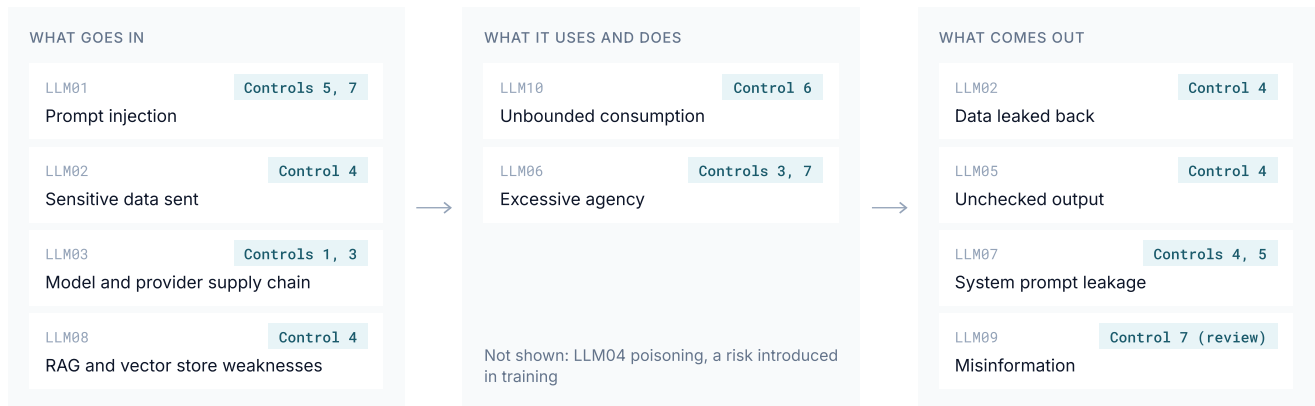
Why existing tools miss it

Data loss prevention was built for files and email, and rarely sees server-to-provider API calls. Firewalls and API gateways see the traffic but not the meaning of a prompt. Provider guardrails cover one vendor only, and a quick internal proxy still leaves redaction, injection detection, key rotation, and audit reporting to build. What is missing is one layer that understands AI traffic, sits inline for every provider, and produces evidence by default.

3. The LLM threat landscape

Most LLM risk enters at three points: what goes into the model, what comes out of it, and what it is allowed to consume or do.

FIGURE 02

LLM risk enters at three points in every request

Badges name the controls from chapter 4 that reduce each risk

Figure 2 · OWASP Top 10 for LLM Applications, grouped by where each risk enters

The figure groups the OWASP Top 10 for LLM Applications by where each risk enters, and the badges show which of the eight controls reduce it. Eight of the ten can be reduced, fully or in part, at a single checkpoint in the request path. MITRE ATLAS adds the attacker's view, helping security teams map each detection to a known technique.

4. The 8 controls for LLM runtime security

Eight controls, enforced at one checkpoint in front of every model, close most of the gap between how fast teams ship AI and how well security can govern it.

They work as a set. The first three decide who may reach which model. The next three protect what flows through. The last two make everything visible and provable.

FIGURE 03

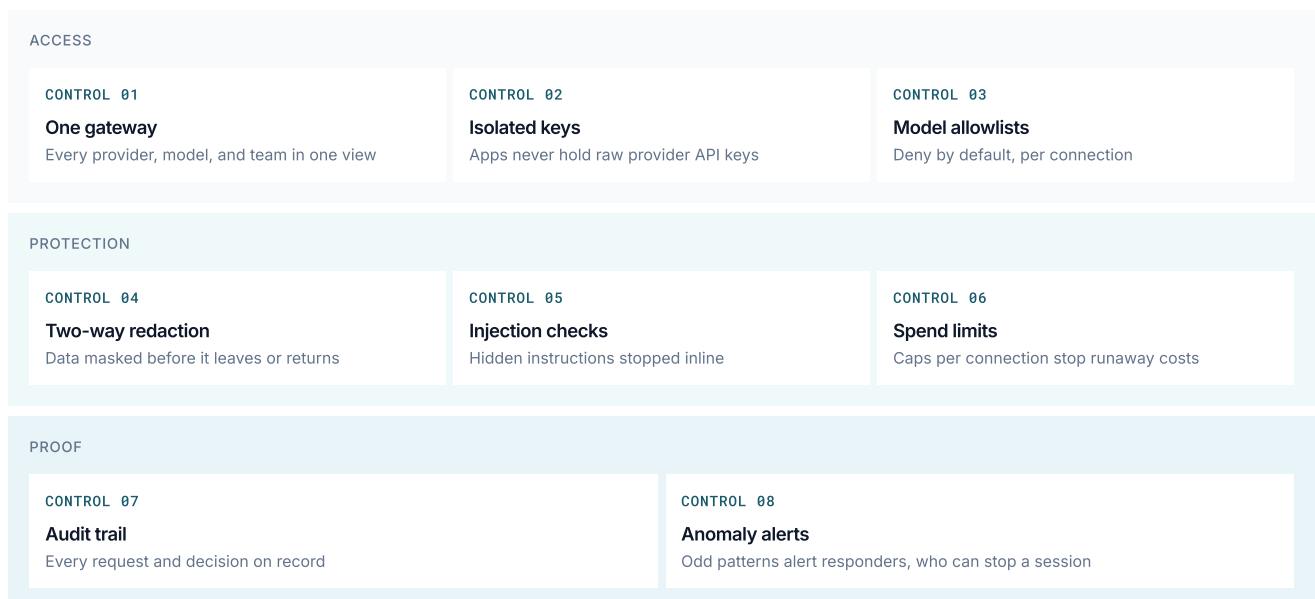
Eight controls: three decide access, three protect the traffic, two prove it

Figure 3 · The eight controls at a glance

FIGURE 04

Every request passes one checkpoint on the way in and on the way out

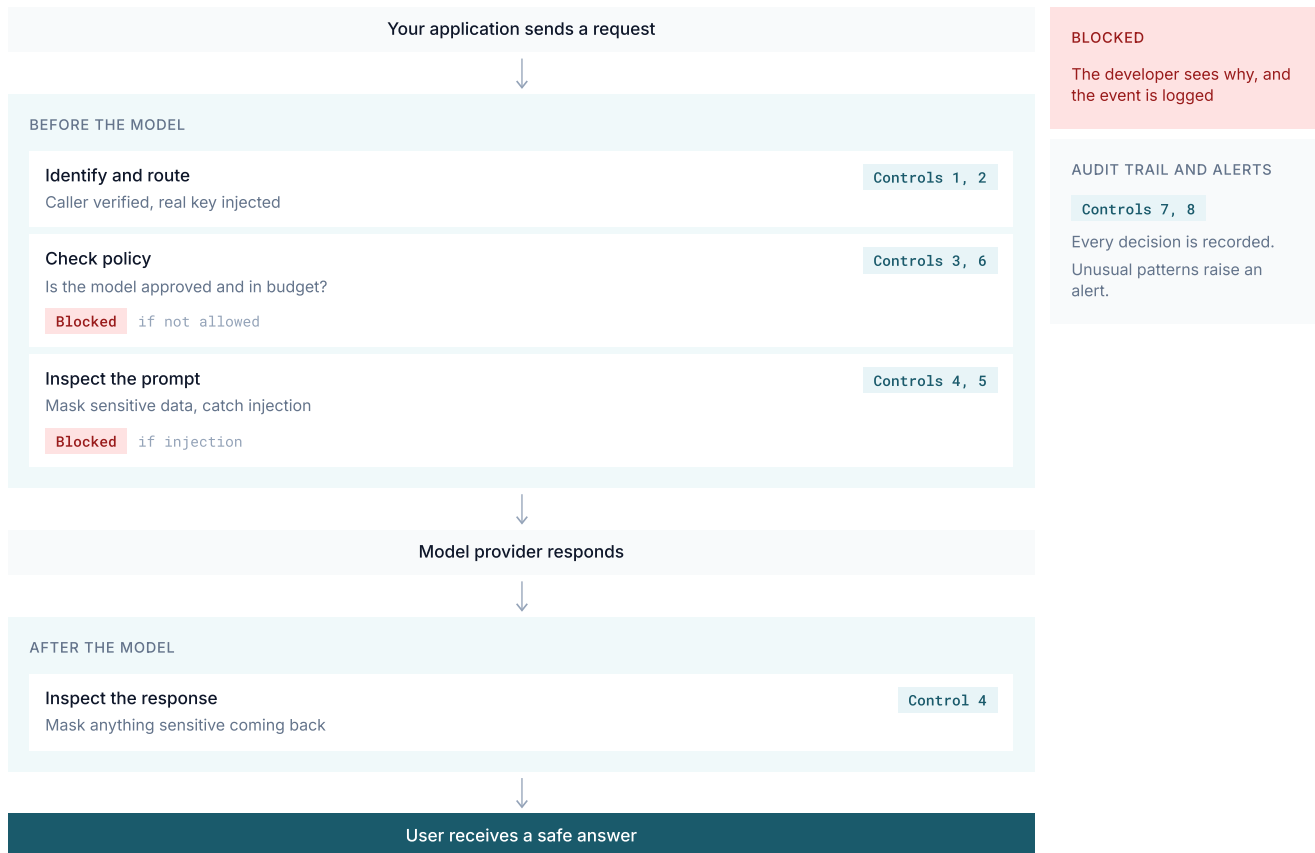


Figure 4 · The life of one LLM request through the eight controls

A request is checked three times before it reaches the model and once more before the answer reaches the user, and every decision lands in the audit trail.

5. Framework crosswalk

Dozens of AI frameworks exist, but they keep asking for the same five things: know your AI, control who uses it, protect the data, keep records, and respond fast when something goes wrong. Build those once, and the same evidence answers most audits.

The table maps each control to the clauses it supports across NIST AI RMF and the Cyber AI Profile, ISO/IEC 42001, 27001 and 27701, ETSI EN 304 223, the EU AI Act, GDPR, DORA, India's DPDP Act, AI Governance Guidelines and CERT-In Directions, plus SOC 2, HIPAA, PCI DSS, and GLBA. No tool makes you compliant on its own. The right layer supplies the controls, and the proof that they work.

CONTROL	STANDARDS (NIST, ISO, ETSI)	EU (AI ACT, GDPR, DORA)	INDIA (DPDP, AI GUIDELINES, CERT-IN)	INDUSTRY (SOC 2, HIPAA, PCI DSS, GLBA)
1. Gateway and inventory	NIST AI RMF Govern (AI inventory); ISO/IEC 42001 operational planning; ISO/IEC 27001 A.5.9 asset inventory	AI Act Art. 26 deployer duties; DORA register of ICT third-party providers	DPDP accountability for processors; Accountability sutra	SOC 2 CC6.1 logical access
2. Credential isolation	NIST Cyber AI Profile Protect; ISO/IEC 27001 A.5.17 authentication information and A.8.24 cryptography	GDPR Art. 32 security of processing	DPDP reasonable security safeguards	SOC 2 CC6.1; PCI DSS Req. 8 authentication

CONTROL	STANDARDS (NIST, ISO, ETSI)	EU (AI ACT, GDPR, DORA)	INDIA (DPDP, AI GUIDELINES, CERT-IN)	INDUSTRY (SOC 2, HIPAA, PCI DSS, GLBA)
3. Model allowlists	NIST AI RMF Manage; ETSI EN 304 223 secure deployment	AI Act Art. 26 use as intended; GDPR Art. 28 approved processors	DPDP processing only through engaged processors; Safety sutra	HIPAA business associate safeguards; SOC 2 CC6.1
4. Two-way redaction	NIST AI RMF Measure (privacy); ISO/IEC 27001 A.8.12 data leakage prevention; ISO/IEC 27701	GDPR Art. 5(1)(c) minimization and Art. 25 privacy by design	DPDP purpose limitation and safeguards	PCI DSS Req. 3 account data; HIPAA minimum necessary; GLBA safeguards
5. Injection detection	NIST AI RMF Measure and Manage; ETSI EN 304 223; NIST Cyber AI Profile Detect	AI Act Art. 15 cybersecurity; GDPR Art. 32	Safety, Resilience and Sustainability sutra	SOC 2 CC7.2 monitoring for anomalies
6. Budget and rate limits	NIST AI RMF Manage; ISO/IEC 42001 resources	DORA ICT risk management	Safety, Resilience and Sustainability sutra	SOC 2 A1.1 capacity management
7. Audit trail	NIST AI RMF Govern and Measure; ISO/IEC 42001 Clause 9 performance evaluation; ISO/IEC 27001 A.8.15 logging	AI Act Art. 12 record keeping and Art. 26 log retention; GDPR Art. 30 records of processing	CERT-In log retention; Accountability sutra	SOC 2 CC7.2; HIPAA audit controls 164.312(b); PCI DSS Req. 10 logging
8. Anomaly detection and alerting	NIST Cyber AI Profile Detect and Respond; ISO/IEC 27001 A.8.16 monitoring	AI Act Art. 14 human oversight; GDPR Art. 33 breach notification; DORA incident management	CERT-In incident reporting; DPDP breach intimation	SOC 2 CC7.3 incident evaluation; HIPAA security incident procedures

A note on the EU AI Act. Articles 12, 14, and 15 apply in law to high-risk AI systems. For every other LLM use, they remain the clearest public benchmark of what responsible logging, oversight, and security look like, and enterprise buyers increasingly use them that way.

6. The LLM Security Maturity Model

Most teams shipping their first AI features start at level one or two, and moving up a single level removes the risks that keep security leaders awake.

FIGURE 05
Each maturity level builds on the one below it

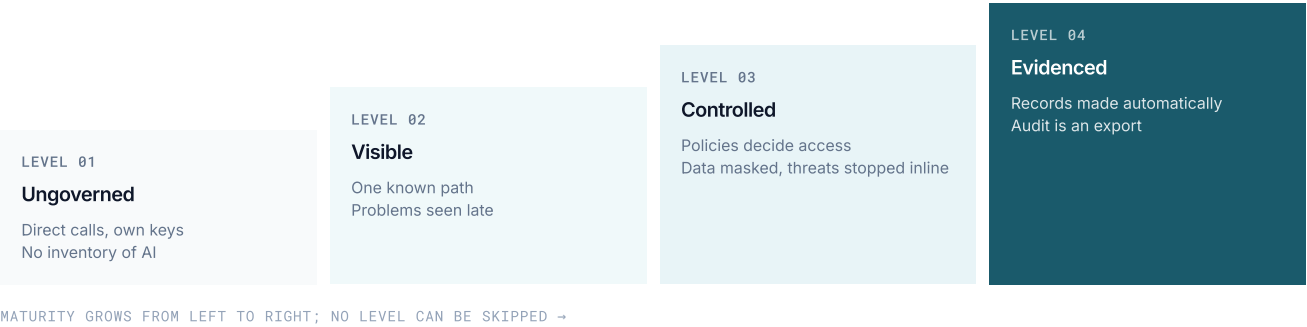


Figure 5 · The LLM Security Maturity Model

Each step up removes a class of risk: level two ends blind spots, level three stops harm inline, and level four turns every control into proof.

Score yourself

Tick each statement that is true for your organization today.

☐ We know every LLM provider and model in use.

☐ No application holds a raw provider API key.

☐ Sensitive data is masked before it reaches a model.

☐ Injection attempts and runaway costs are stopped inline.

☐ Every request and policy decision is logged.

☐ We can export audit evidence on request.

How to read your score. Zero or one tick: Level 1, Ungoverned. Two or three: Level 2, Visible. Four or five: Level 3, Controlled. All six: Level 4, Evidenced.

7. Industry playbooks

The eight controls apply everywhere, but each industry has one risk that matters more than the rest, and that risk decides where to start.

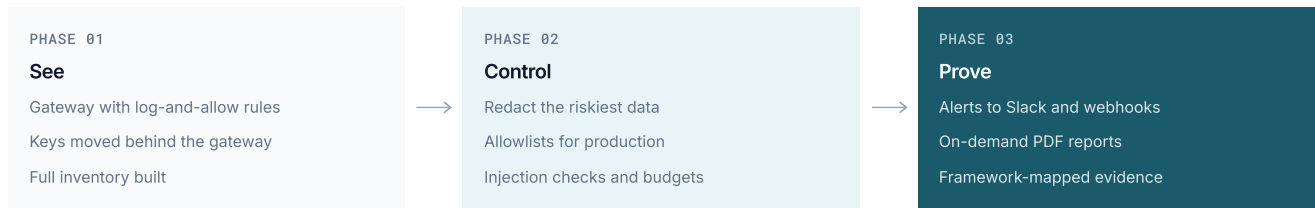
INDUSTRY	THE RISK THAT MATTERS MOST	KEY FRAMEWORKS	START WITH
Fintech and banking tech	Card, account, and identity data reaching a provider outside the compliance boundary	PCI DSS, GLBA, DORA, GDPR, DPDP Act	Controls 4, 7, then 3
Healthtech	Patient data sent to a service without the right agreements or safeguards	HIPAA, GDPR, DPDP Act, ISO/IEC 27001	Controls 4, 3, then 7
Legal tech	Confidential client material reaching an unapproved model	GDPR, DPDP Act, ISO/IEC 27001, SOC 2	Controls 3, 4, then 7
Enterprise SaaS	Enterprise deals stalling on "where does our data go?"	SOC 2, ISO/IEC 27001 and 42001, GDPR, EU AI Act	Controls 1 and 7 together

8. A three-phase rollout

The fastest path to secure AI is not a big-bang project but three phases, each delivering value on its own: see everything, control everything, then prove everything.

FIGURE 06

Three phases, each delivering value on its own



Start with visibility to earn trust; controls then land smoothly

Figure 6 · The three-phase rollout

Each phase stands on its own: stop after Phase 1 and you still have the first complete picture of your AI traffic.

9. Buyer's checklist

Before choosing any LLM security platform, ask these eight questions; a strong vendor answers each one with a demo, not a slide.

1. Does it cover every model provider we use, and new ones as we add them?
2. Can developers adopt it through configuration, without rewriting code?
3. Are provider keys encrypted, and can we rotate them without downtime?
4. Is it deny-by-default, so nothing reaches a model without a policy?
5. Does it inspect responses as well as prompts, including our own data patterns?
6. How does it detect prompt injection, and what happens when it does?
7. Can a responder see every request and stop a session immediately?
8. Can it export evidence mapped to our frameworks, and can we retain it for as long as our auditors need?

10. How Igris delivers this

Igris puts all eight controls into one platform that sits between your applications and every model provider, and your developers adopt it through configuration rather than code rewrites.

Two Igris products carry this paper's controls. **Igris Guard** is a transparent proxy between your users and any LLM, built on a unified gateway with one OpenAI-compatible API across 50+ chat providers, plus any OpenAI-compatible or self-hosted endpoint such as Ollama. **Igris Lens** streams every event into one queryable timeline, with real-time alerts, risk heatmaps, and audit-ready reports.

FIGURE 07

Igris sits between every application and every model, and records it all

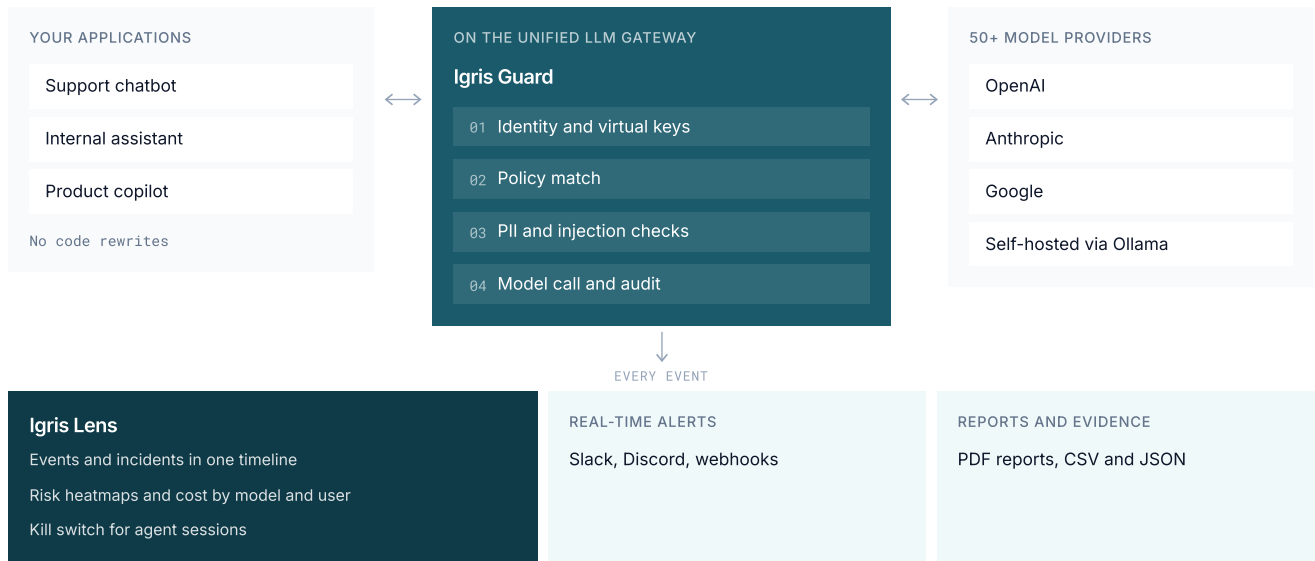


Figure 7 · Where Igris Guard and Igris Lens sit in your AI stack

Guard enforces policy on every request in both directions, and Lens turns every decision into alerts, reports, and audit evidence.

Every control, mapped

CONTROL	HOW IGRIS DELIVERS IT
1. Gateway and inventory	One gateway for 50+ chat providers, with every connection, model, and user visible in Lens
2. Credential isolation	Virtual keys hide provider credentials, which are encrypted at rest with AES-256-GCM and rotate with a single API call and no downtime
3. Model allowlists	Every new connection starts deny-by-default; each connection (virtual key) is restricted to approved models, and blocked requests return a clear reason
4. Two-way redaction	Built-in detectors for email, US phone numbers, US SSN and UK National Insurance numbers, Luhn-validated card numbers, cloud and platform secrets (AWS, GCP, GitHub, Slack, JWTs, private keys), and IPv4 addresses, plus your own regex and keyword patterns, applied to prompts and to non-streaming responses
5. Injection detection	Keyword-based prompt injection detection, with four policy actions: allow, alert (log and allow), redact, or block
6. Budget and rate limits	Request, token, and daily spend limits per connection, with cost broken down by model, connection, and user in Lens
7. Audit trail	Every event captured with user, model, decision, and trace ID; filter by user or trace ID, export as CSV, and pull JSON evidence through the SOC 2 export
8. Anomaly detection and alerting	Related agent events roll up into incidents with severity, alerts reach Slack, Discord, or any HTTP webhook, and any MCP agent session can be suspended in one click

Developers adopt Igris through SDK adapters for OpenAI and Anthropic, with Google models reachable through the OpenAI-compatible API, so existing code keeps working. Security leaders get one risk dashboard and PDF compliance reports, and every proxied request is captured in the audit trail.

See it on your own traffic

The best way to judge any of this is on your own AI traffic. Book a demo at igrisecurity.com and we will walk through the eight controls live, starting with Phase 1: see everything.

Sources

- **Standards and frameworks:** NIST AI Risk Management Framework, NIST AI 600-1, NIST IR 8596 Cyber AI Profile (preliminary draft), ISO/IEC 42001, ISO/IEC 27001, ISO/IEC 27701, ETSI EN 304 223
- **Threat references:** OWASP Top 10 for LLM Applications, MITRE ATLAS
- **European Union:** EU AI Act, GDPR, DORA
- **India:** DPDP Rules (obligations phasing in through 2027), India AI Governance Guidelines, CERT-In Directions
- **Industry and assurance:** SOC 2 Trust Services Criteria, HIPAA Security Rule, PCI DSS, GLBA Safeguards