BUILT BY A TEAM THAT HAS SECURED MAJOR INSTITUTIONS WORLDWIDE FOR 25 YEARS.

ARGENT RESPONSE SCREENING

MOST TEAMS GUARD WHAT GOES OUT. ARGENT CHECKS WHAT COMES BACK.

Argent Strategic Intelligence sits between your AI agents and the models they use. A careful prompt can still bring back an answer with a customer's Social Security number in it. In Enforce, Argent holds every reply that comes back through it until the check is done. OWASP ranks sensitive information disclosure through model output a top AI app risk.[1]

You choose what a hit does. Redact, with the default Redact + Audit policy, blanks the sensitive spans and delivers the rest of the answer. Block withholds the whole reply and names the kinds of data found. Every decision goes on a signed record as categories and counts.



SEE.
What each model sent back, reply by reply.



PROVE.
A signed record of what was found and done.



STOP.
Redact or Block: your admin sets what a hit does.

WHAT ARGENT CHECKS ON THE WAY BACK

1	**Sensitive data** A model's account summary repeats a customer's SSN.	The detector and DLP policy that screen prompts read what the model wrote, including tool calls. With Redact and the Redact + Audit policy, the spans are blanked and the rest goes through.
2	**Block mode** The bank wants no flagged reply delivered at all.	An admin sets Egress leak handling to Block. In Enforce, every reply with a detection is withheld; the agent gets a 403 naming the data types and a trace id, never the value.
3	**Hidden image tricks** A reply hides an image tag that would send data out.	In Enforce, external markdown images and disguised links in the model's text are replaced with a fixed marker such as [image blocked by Argent].
4	**Classifier and rules** A rule bars pending deal terms.	The optional Classifier can score the reply; your plain-language rules run on it too. A rule set to Block withholds a match; Warn raises an alert. A Classifier hit is recorded as flagged.
5	**Fail closed** The detector cannot run.	In Enforce, by default, a reply that cannot be checked is not released: the agent gets a 403 saying the response was withheld under egress policy, with a trace id.

WHERE THE PRESSURE IS WRITTEN DOWN

NY DFS, 23 NYCRR 500.17 (covered entities). Notice of a cybersecurity incident "in no event later than 72 hours" after determining it occurred.[2]

HIPAA, 45 CFR 164.312(b) (covered entities and business associates). Mechanisms that "record and examine activity" in systems holding health data.[3]

HOW A TEAM STARTS

1. Watch. Argent records what it would have done. Every reply is delivered unchanged.

2. Enforce. Switch the DLP area to Enforce. Under the default policy, Redact blanks spans; choose Block where nothing flagged may pass.

3. Review. The signed record shows, reply by reply, what was found and what Argent did, as categories and counts.

BOOK A DEMO. SEE WHAT YOUR AI MODELS SEND BACK, AND DECIDE WHAT GETS THROUGH.

NYEDIS.AI/DEMO

Sources. 1 OWASP GenAI Security Project, LLM02:2025 Sensitive Information Disclosure, genai.owasp.org. 2 23 NYCRR 500.17(a)(1), Notices to superintendent (Cornell LII). 3 45 CFR 164.312(b), HIPAA Security Rule technical safeguards, audit controls (Cornell LII).