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

ARGENT PROMPT INJECTION DEFENSE

YOUR AGENTS READ TEXT YOU NEVER WROTE. ARGENT MARKS IT AS DATA.

Argent Strategic Intelligence sits between your AI agents and the models they use. A web page, document or reply an agent fetches through a tool can hide a line that gives it orders. Argent wraps that text in a marker naming its source and tells the model it is data. OWASP ranks prompt injection first in its 2025 Top 10 for AI apps.[1]

A person can try it too. When a prompt tries to override the agent's instructions, Argent's Classifier, or your own guardrail, scores it and refuses a hit before the model sees it. Every decision goes on a signed record, never the text itself.

SEE.
Every refused prompt and repeat offender, by agent.

PROVE.
A signed record of each decision, as codes, counts and scores.

STOP.
Override prompts are refused before the model.

FIVE LAYERS IN ONE CHECKPOINT

1	**Untrusted Fencing** A vendor web page tells the agent to send out a loan file.	Tool results and agent replies are wrapped in a random marker with a source tag, and the model is told it is data. A forged end marker is defaced.
2	**Classifier** A prompt says "ignore your instructions and..."	Scores the person's latest prompt for prompt injection and jailbreak against a threshold you set, and refuses a hit with the code classifier.prompt_injection.
3	**Your own guardrail** The security team already runs AWS Bedrock Guardrails.	Route the decision to your own guardrail, AWS Bedrock Guardrails or LLM Guard, and Argent enforces its verdict. If the vendor does not answer, the request is refused.
4	**Plain-language rules** Compliance wants pending deal terms kept out of AI.	Write a rule in plain words about a topic or kind of data, then choose Shadow, Warn or Block.
5	**Detections** One agent keeps tripping the refusal.	Five prompt-injection refusals by one agent in one ten-minute window raise an "injection block repeat" detection and raise that agent's threat score.

WHAT SECURITY AGENCIES ADVISE

UK NCSC, December 2025. AI models "do not enforce a security boundary between instructions and data". It advises marking data apart and logging enough to spot abuse.[2]

NIST AI 100-2e2025. Indirect prompt injection comes from a third party through content the AI reads, and can hijack an agent.[3]

HOW A TEAM STARTS

1. Watch. The Classifier's Observe tab shows what it would flag, blocking nothing.

2. Enforce. Switch the Classifier to Sync Block; move rules to Block.

3. Review. Detections and threat scores show which agents keep drawing prompt-injection refusals.

BOOK A DEMO. SEE HOW ARGENT TREATS THE TEXT YOUR OWN AGENTS READ.

NYEDIS.AI/DEMO

Sources. 1 OWASP GenAI Security Project, LLM01:2025 Prompt Injection, genai.owasp.org. 2 UK NCSC, Prompt injection is not SQL injection (it may be worse), 8 Dec 2025. 3 NIST AI 100-2e2025, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, March 2025.