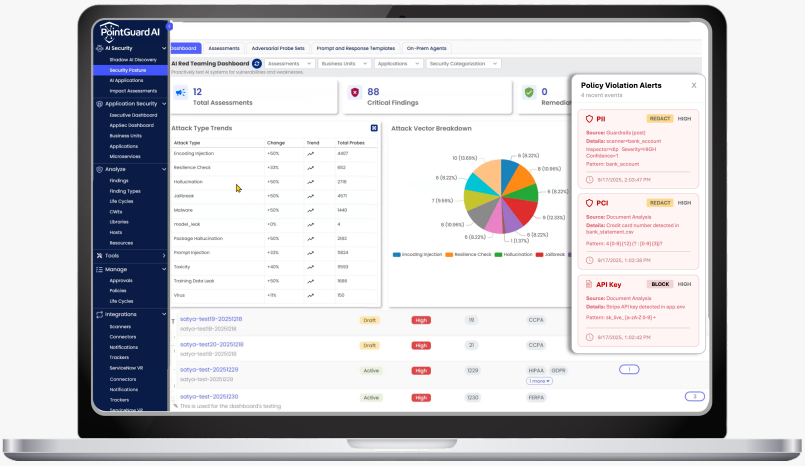
PointGuard AI Launches Advanced Guardrails to Prevent Indirect Prompt Injection Attacks

New protections inspect documents, metadata, prompts, and responses before AI models can be manipulated



PointGuard AI Advanced Security Guardrails

SAN JOSE, CA, UNITED STATES, March 4, 2026 /EINPresswire.com/ -- PointGuard AI today announced the availability of Advanced Guardrails designed to prevent Indirect Prompt Injection attacks, one of the fastest-growing security threats in enterprise AI environments.

As organizations deploy AI agents that retrieve documents, query databases, and connect to external tools via MCP and APIs, a new attack vector has emerged. Unlike conventional prompt injection that directly manipulates a user's query, indirect prompt injection embeds malicious instructions inside documents, source code, PDFs, or metadata. When those files are later retrieved and passed to a large language model for processing, hidden instructions can hijack the model, trigger unauthorized tool execution, or exfiltrate sensitive data.

> Indirect prompt injection has become a serious risk because the attack surface has shifted from the prompt to the data itself. Our Advanced Guardrails provide that critical enforcement layer."
>
> *Pravin Kothari, CEO of PointGuard AI*

PointGuard's Advanced DLP Guardrails address this threat through its secure gateway architecture. Acting as a control layer between AI agents, tools, and models, PointGuard continuously monitors prompts, responses, and attached documents before they reach the LLM.

The solution delivers:

- Inspection of prompts and model responses to detect prompt injection, jailbreak attempts, and policy violations
- Deep scanning of retrieved files, including source code and document metadata, to identify embedded malicious prompts
- Advanced DLP controls that redact PII, secrets, API keys, and sensitive corporate data before

model ingestion
- Real-time blocking of malicious instructions that attempt unauthorized tool execution

As demonstrated in customer environments, when an AI agent retrieves a file from GitHub or another MCP-connected tool, PointGuard inspects the content before it is sent to the model. If hidden prompt injection text or malicious metadata is detected, the gateway blocks the content and prevents downstream compromise. This applies equally to externally sourced documents and internally stored files that may have been compromised earlier in the supply chain.

PointGuard's guardrails operate with minimal performance impact by leveraging optimized detection models designed specifically for prompt injection and DLP enforcement. The architecture supports hybrid deployment models, allowing inspection modules to run locally within customer environments to preserve data sovereignty.

"Indirect prompt injection has become a serious enterprise risk because the attack surface has shifted from the prompt to the data itself," said Pravin Kothari, CEO of PointGuard AI. "Security controls must now inspect not only what users type, but also what agents retrieve and what models consume. Our Advanced Guardrails provide that critical enforcement layer."

For more information, visit www.pointguardai.com.

Willy Leichter
PointGuard AI
email us here
Visit us on social media:
LinkedIn
YouTube