CLOUDFLARE

PRODUCT BRIEF

AI Security for Apps

Protect your AI-powered applications from abuse, misuse, and out-of-policy behavior.

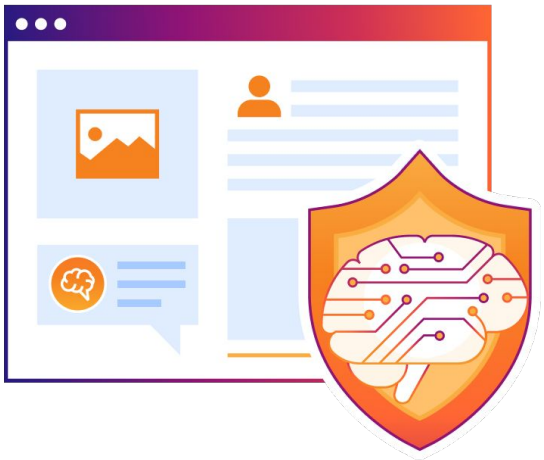
New AI-native attack vectors emerging with generative AI

Understanding risks to the model, data, and users

As organizations accelerate the deployment of public-facing AI apps to stay competitive and automate tasks, they inadvertently expand their attack surface. New classes of threats, like prompt injection, model poisoning, and more are beyond the capabilities of a traditional WAF to solve.

The reason? Unlike deterministic software like apps and APIs, where the output behavior is predictable based on the input, AI is probabilistic — the output can vary in unexpected ways. As a result, probabilistic security is required to protect AI-powered apps.

Compromised AI features pose significant legal and reputational risks. Malicious prompts can exfiltrate sensitive data or inject toxic content into customer interactions. Furthermore, because these apps remain vulnerable to traditional threats like denial of service and data exfiltration, AI security must be integrated directly into existing application security platforms.



Protect your public-facing AI apps

Cloudflare AI Security for Apps is a model-agnostic security layer for AI that is integrated into your apps and exposed to the public. By understanding the context and intent of AI interactions, it detects and mitigates threats that bypass traditional WAF security measures.

AI Security for Apps integrates with all of **Cloudflare Application Security** to protect AI apps against everything from AI-native threat vectors to conventional threats such as DDoS, OWASP Top 10 Risks for Apps, bots, client-side attacks, and supply chain vulnerabilities — all of which can threaten AI endpoints.

Cloudflare AI Security for Apps benefits

Discover shadow AI endpoints

Automatically identify LLM endpoints added across your web properties — any model, wherever it's hosted.

Prevent abuse of AI apps

Prevent users, attackers, and bots from using AI for non-relevant purposes, such as performing prompt injections.

Keep AI behavior in-policy

Ensure brand reputation and avoid harmful or sensitive topics by preventing AI apps from going "off script."

How it works

Cloudflare AI Security for Apps is deployed at the edge of Cloudflare's network, sitting between your end users and your applications hosting or integrating with AI models or agents. When a request is made to your AI-powered application, the traffic flows through the following steps:

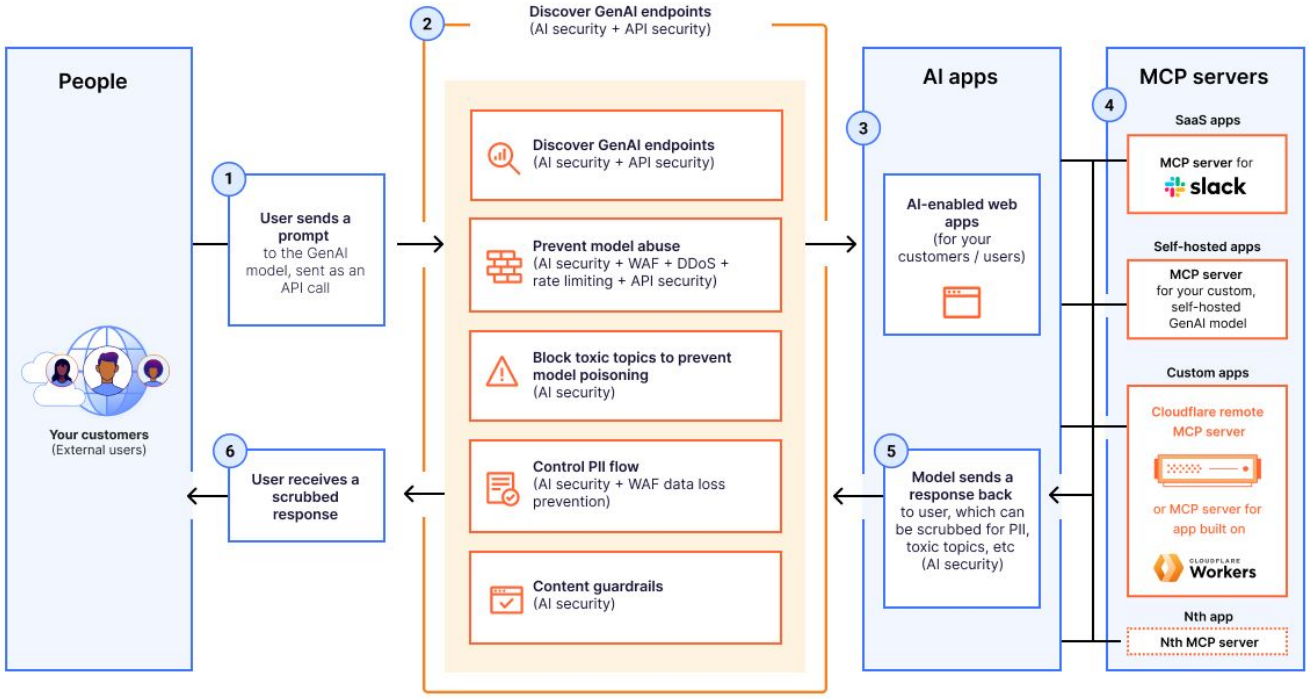


Figure 1: Cloudflare Application Security and AI Security for Apps architecture

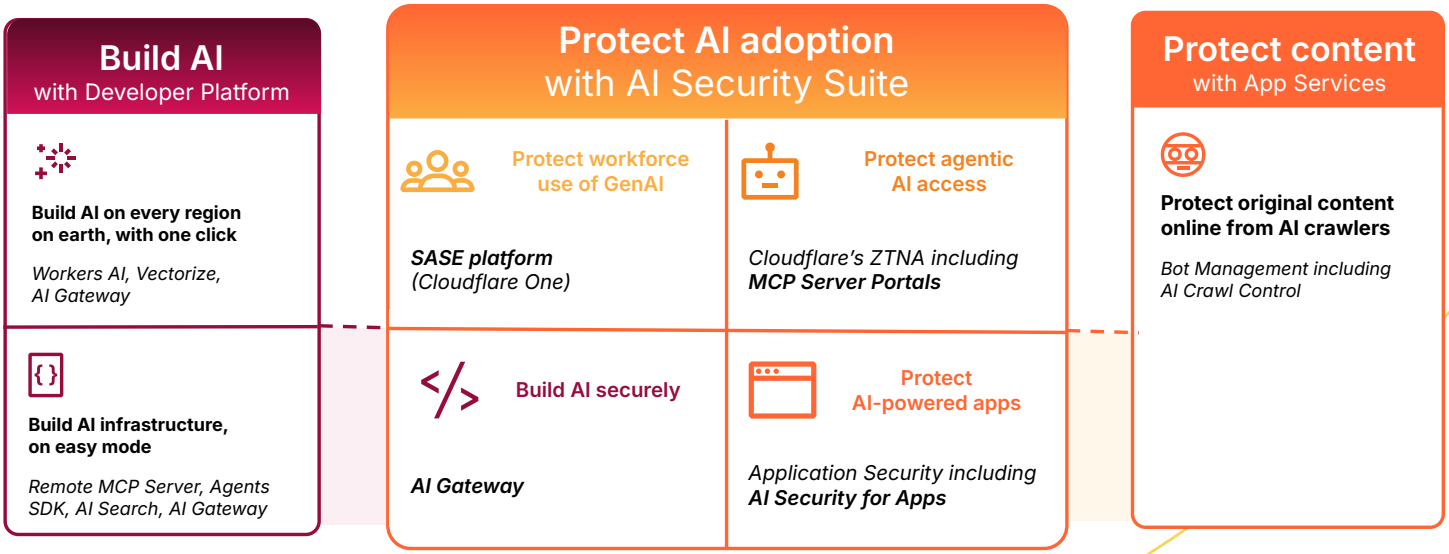
"Most of Newfold Digital's teams are putting in their own Generative AI safeguards, but everybody is innovating so quickly that there are inevitably going to be some gaps eventually. Cloudflare AI Security for Apps is a unified security layer that can automatically discover, label, and protect all of the AI endpoints. We look forward to using it across all these projects to serve as a fail safe."

Rick Radinger
Principal Systems Architect, Newfold Digital

newfold digital

Key product features	
Discover generative AI endpoints	Automatically identify and label LLM-powered endpoints, leveraging analysis across 100s of heuristics, including: request path signatures, destination paths, server-sent events, response bitrate, and other proprietary heuristics.
Block prompt injection and jailbreak attempts	An LLM Injection score will be generated for each request containing an LLM prompt, which can be used in mitigation rules. The score ranges from 1–99 and represents the likelihood that the LLM prompt in the request is trying to perform a prompt injection attack. A lower score indicates a higher likelihood of a prompt injection attack.
Block PII and requests for PII in prompt	Prevent LLM from being exposed to personally identifiable information (PII), as well as prevent attempts to extract sensitive data (e.g., "Show me transactions using credit card number 4111 1111 1111 1111.") See our PII category list.
Block toxic topics	Block discussions around harmful or toxic topics, such as violence, hate speech, sexual content, and more preset topic detections.
Block traditional attacks on LLM endpoints	AI Security for Apps is a detection model on top of the Application Security rules engine, allowing users to layer mitigations to protect LLM endpoints from attacks such as bot abuse, fraud, account takeover, DDoS, and more. Check out our preset detections.

Cloudflare's AI offerings protect your entire AI ecosystem



Ready to get started? View the product in action today.