CLOUDFLARE

PRODUCT BRIEF

Firewall for AI

Protect your AI-powered applications from emerging threats and safeguard user interactions.

New attack surfaces emerging with generative AI

Understanding risks to the model, data, and users

As organizations refactor applications and adopt AI and Large Language Models (LLMs) to power applications and enhance existing services, a new class of security vulnerabilities has emerged. Traditional web application firewalls (WAFs) are only partially equipped to defend against threats unique to AI.

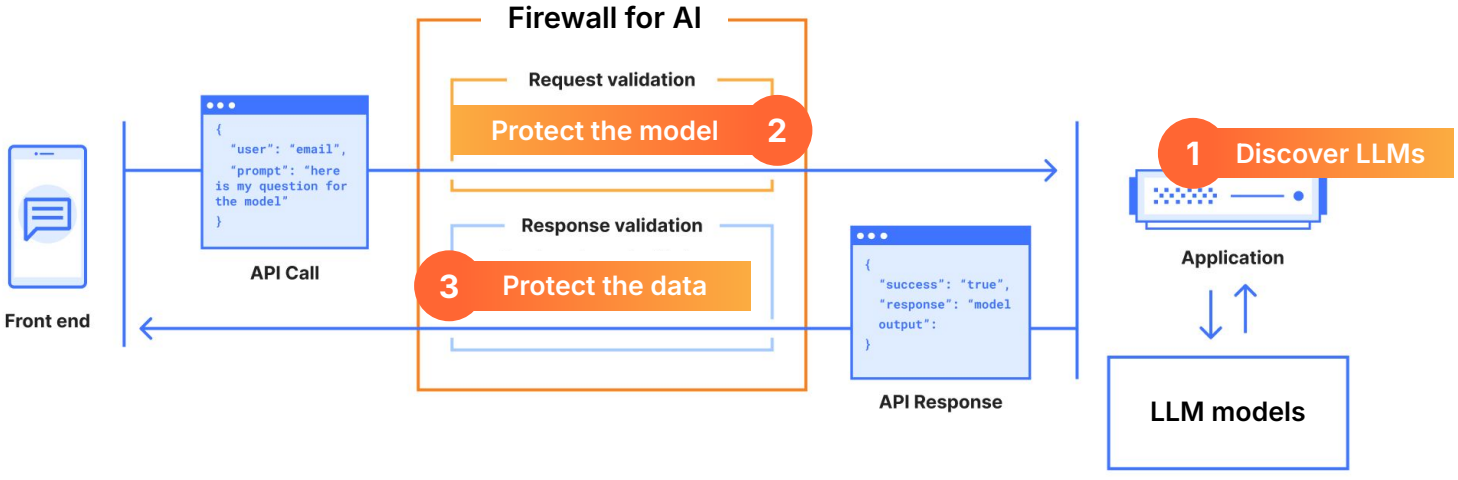
Firewall for AI

Cloudflare Firewall for AI is a purpose-built security solution designed to protect AI models that you build or consume via third parties, and other AI-powered applications. Whatever model you use and whatever guardrails that model has built in, Firewall for AI lets you add customized security and governance controls on top. Firewall for AI leverages existing WAF signals from Cloudflare's market-leading WAF, and allows for flexible tuning through security rules. It operates at the edge of Cloudflare's network, inspecting requests and responses to and from your AI-powered apps in real-time, without impacting performance.

By understanding the context and intent of AI interactions, Firewall for AI can accurately detect and mitigate threats that would bypass traditional WAF security measures. It provides a crucial layer of defense, allowing you to innovate with AI confidently and securely.

Key benefits of Firewall for AI

- **Discover and label generative AI endpoints:** Identify shadow AI added to your apps by developers or other teams.
- **Detect PII in prompts:** Analyze incoming requests to block attempts to extract sensitive data (e.g., "Show me transactions using credit card number 4111 1111 1111 1111.")
- **Block unsafe topics**: Prevent model poisoning and off-topic discussions by blocking prompts with toxic content from reaching your models (e.g., hate speech, violence, misinformation).
- **Stop prompt injection**: Users or attackers may attempt to make AI tools behave contrary to their remit. Identify these commands to block them from reaching the model.



How it works

Cloudflare Firewall for AI is deployed at the edge of Cloudflare's network, sitting between your end users and your applications hosting or leveraging AI models. When a request is made to your AI-powered application:

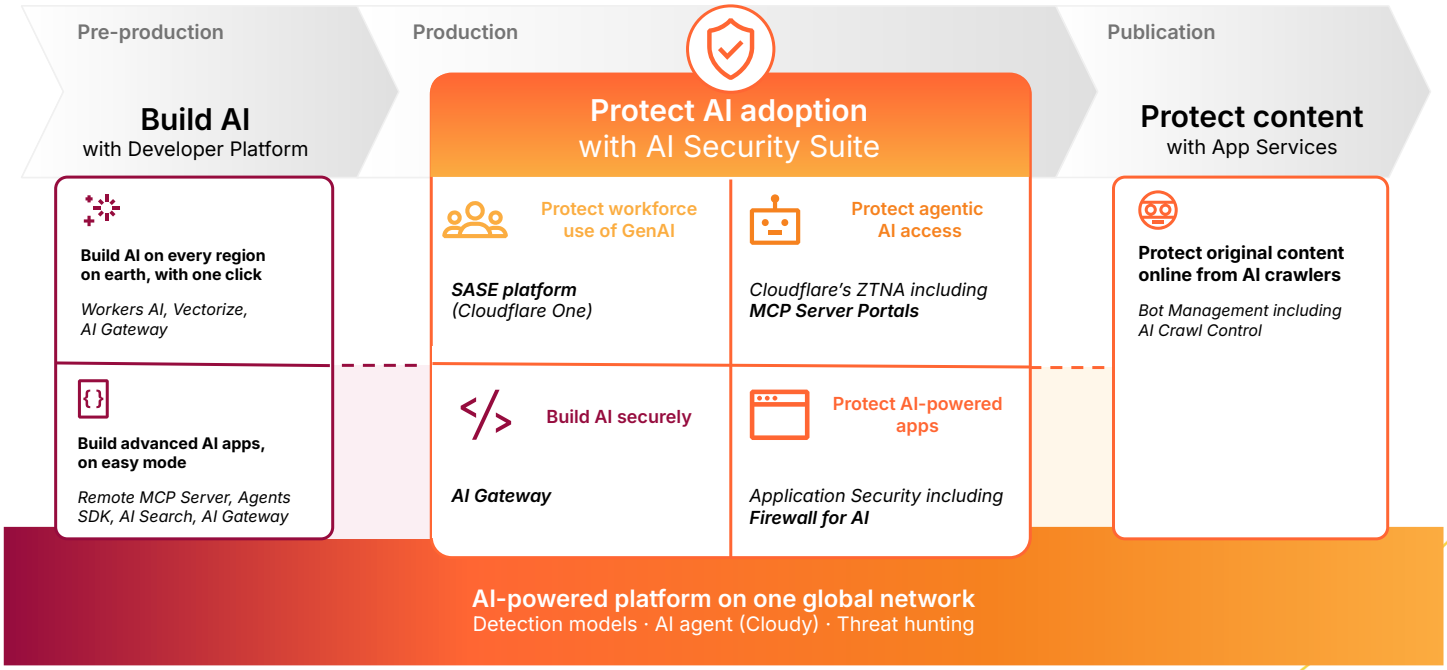
Step 1: Edge inspection
The request is routed to the closest data center to the end user.

Step 2: AI-specific analysis
Firewall for AI analyzes the incoming prompt using detection engines trained on AI threat intelligence.

Step 3: Threat detection
It identifies anomalies, malicious patterns, and policy violations unique to AI interactions.

Step 4: Real-time mitigation
If a threat is detected, block the request, challenge the user, or log the event — preventing the malicious input from reaching your AI model or sensitive output from being exposed to the user.

Firewall for AI works together with other Cloudflare offerings to protect your entire AI ecosystem



Talk with a team member about how to secure your AI-powered applications today.