

Fastly for AI

Govern. Protect. Control.

One platform to build, secure, and deliver your apps, AI, and data inline - in single-digit milliseconds, on the network already carrying your traffic.

The bottom half of the image features a dark background with a series of thin, white, wavy lines that create a sense of motion and depth. These lines curve from the bottom right towards the top right, with a color gradient transitioning from dark blue on the left to a vibrant red on the right. In the bottom left corner, the Fastly logo is displayed in white.

fastly®

Why now?

Security and cost concerns are growing as AI use expands. 80% of organizations have already encountered risky behavior (like unauthorized actions, data exposure, and operational errors that happen outside of set safety rules) from AI agents.* And 93% of organizations are exceeding their AI budgets.**

The resources to mitigate these risks posed by AI are keeping security, engineering, and business leaders chronically behind. Gaining control over AI starts with understanding what AI is doing across the business – a task only possible with a solution purpose-built to govern, protect and control AI.

Fastly for AI

Govern. Control. Protect.

All businesses are striving to move faster with AI. To do so safely, organizations require deep insight into AI behavior, and most don't have a complete picture of what their AI is actually doing across the business.

Leaders feel they're at the mercy of uncontrolled AI: Spend arrives as a surprise invoice, prompts and completions go unlogged, and new AI features introduce a growing attack surface. And all of this lands on APIs never designed for this type of caller.

Fastly set out to resolve this control gap. Fastly already sits between your users and your applications, and we are now extending that to AI. The Fastly platform delivers a single control plane for every AI call, AI token usage oversight, enforcement on the APIs AI reaches for, and in-line defense for the AI features you ship.

6.5x

AI traffic is growing 6.5x faster than human traffic and is approaching half of all requests.

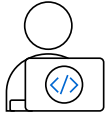
[AI Runtime Control](#) | [API Security](#) | [AI Firewall](#) | [AI Accelerator](#) | [AI Bot Management](#) | [MCP Server](#)

*Source: McKinsey Global Survey on AI, 2025; McKinsey Quarterly, 2025.

**Source: "Burning through the AI budget", July 30, 2026; McKinsey

While the fundamentals of your business haven't changed, AI is adding complexity.

AI touches everything across the modern business, and is changing the way you build with AI, what you power with AI, and how AI interacts with your sites and applications. Businesses must address all three avenues in order to take control of AI within their business ecosystem.



What you build with AI

Engineering teams and the functions around them were told to augment their work with AI, and that mandate arrived without a budget control attached. Every request to a model costs tokens, and in most organizations nobody can attribute that spend to a person, a team, or a project.

What keeps you up: Spend with no ceiling. Unlogged prompts and completions. Credentials spread with no attribution.



What you power with AI

Chatbots, copilots, AI-driven search, recommendations, and support tools are now embedded in the products your business runs. Shipping these features introduces an attack surface existing controls were not built for.

What keeps you up: An attack that arrives as ordinary content and executes elsewhere. Answers nobody can reconstruct.



What AI consumes from you

AI does not browse, it calls. The overwhelming majority of AI agent-to-agent interactions with a business happen through APIs, and increasingly through servers built to let AI query a business directly. An API needs to enforce its contract against a caller that assembles requests at runtime

What keeps you up: Improvising instead of following the contract. Decisions made with no human in the loop.

One platform, many layers of control.

Fastly for AI brings visibility, control, and enforcement to every AI interaction across your business – from the model call an engineer makes to the API contract an agent has to honor. The same platform, the same data path. Nothing new to run.

Get visibility into every AI interaction

Govern

You cannot govern what you cannot see. Fastly captures every AI call your business makes and attributes it to a key, a session, and where SSO is in use, a person. Full prompts and completions are logged, searchable, and filterable. Token counts, latency, and cost land in one place. Month-to-date spend, request volume, and a leaderboard view show who is spending what before the invoice does.

Powered by: [AI Runtime Control \(ARC\)](#)

Control AI use and costs

Control

Enforce API contracts, applying consistent rules to agentic, agent-assisted, and conventional traffic. Fastly lets you issue and rotate virtual keys instead of handing out raw provider credentials. You can set budget limits and alerts, rate limit per key, and failover between providers when one degrades. Fastly even lets you route to your own self-hosted models. In addition, on the API side, you can enforce the specification you already wrote, in logging mode or blocking mode, service by service.

Powered by: [AI Runtime Control \(ARC\)](#) and/or [API Security](#)

Faster AI

ARC and semantic caching boosts performance

Get better AI performance with intelligent caching that understands your data. Fastly's AI Accelerator boosts the performance of popular LLMs like OpenAI and Google Gemini by 9x. No rebuild necessary, just one line of code.

Powered by: [ARC](#) and/or [AI Accelerator](#)

Limit what AI is allowed to do

Protect your AI systems from misuse/abuse

Mitigates LLM attacks, including prompt injection, delivering protection at the edge directly in the request path. AI Firewall evaluates prompts inline, deterministically, with no model in the request path and no GPU hardware. Malicious instructions are identified and blocked before they reach the models they target. This helps cover the attack you can't see coming: instructions that arrive as ordinary content in one place and execute in a downstream model somewhere else.

Fastly Next-Gen WAF accurately detects attacks without tuning using a proprietary SmartParse detection engine with the flexibility to deploy anywhere, to protect everywhere you operate. It deceives attackers to combat their scale and reduce their ability to retool.

Powered by: [AI Firewall](#) (requires ARC) and/or [Next-Gen WAF](#)

Make Bots play by your rules

Protect your bottom line from the machines

Bot Management tells verified AI from traffic merely claiming to be AI, then decides which automation you welcome, rate limit, block, or monetize, because not all AI traffic is bad traffic.

[Fastly ContentGuard](#) is a security feature within Fastly Bot Management that inspects web traffic at the network edge before it reaches your cache or origin server.

Powered by: [Bot Management](#) and ContentGuard

AI-powered edge control

Streamline operations with AI tools

The Fastly Model Context Protocol (MCP) Server and the Fastly Agent Toolkit give your AI agents access to Fastly through the CLI and the skills to work effectively with the Fastly platform. Manage infrastructure, performance, and security through natural language commands.

Powered by: [MCP server](#) and [Fastly Agent Kit](#)

Why Fastly for AI

One data path, not a tool chain

Gateway, API enforcement, and inline inspection sit where your traffic already runs. No extra hops, no extra vendors, no second set of logs.

Deterministic inspection, not a model in the way

AI Firewall uses contextual parsing. Inline, predictable, with no GPU dependency. A solution in every request path is slower, more expensive, and noisier.

Both halves of the AI problem, one platform

Bot Management, DDoS Protection, and Next-Gen WAF handle AI interacting at your business. AI Runtime Control, API Security, and AI Firewall handle AI inside it.

Your providers, your keys, your models and your perimeter

Bring your own provider relationships and keys. Register self-hosted and custom models as first-class backends. Run AI Firewall on premises when prompt payloads cannot leave. One endpoint across hundreds of public and self-hosted models.

Model-agnostic

Fastly does not host or run models and holds no model partnerships, so routing serves you, not us. Your traffic goes where you decide.

Enforcement is the unlock, not the overhead

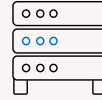
In regulated industries, security and compliance block AI adoption because there is no governance layer to approve against. Visibility and control are not the price of speed. They are the precondition.

Proven at the scale AI demands



622 Tbps

Global edge capacity
(2026-06-30)



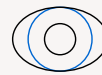
166 POPs

Across 37 countries
(2026-06-30)



5T+

Requests served daily
(2026-03-31)



8T+

Security requests inspected
monthly (2026-03-31)



97%+

Customer satisfaction for five
consecutive years (2026-06-30)



235%

ROI, per [Forrester TEI Study](#)

Fastly already protects you from AI interacting with your business - now we can help you enforce the AI you use internally.

Move Fast. Stay safe.

One platform to build, secure, and deliver your apps, AI, and data.

Learn more at fastly.com/ai-hub