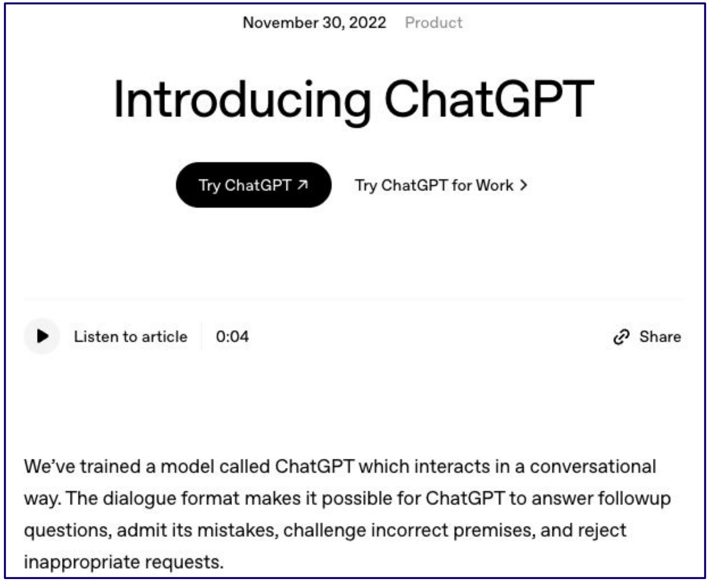
Zero Trust

In the era of Artificial Intelligence

José Dores
Principal Solutions Architect
Cloudflare

AI is the next paradigm shift...

AI is the next paradigm shift...

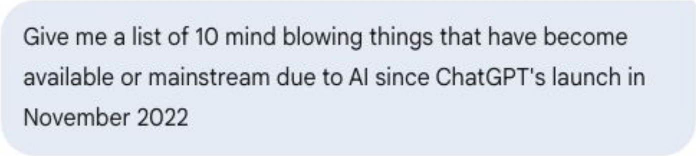


November 30, 2022 Product

Introducing ChatGPT

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

AI is the next paradigm shift...



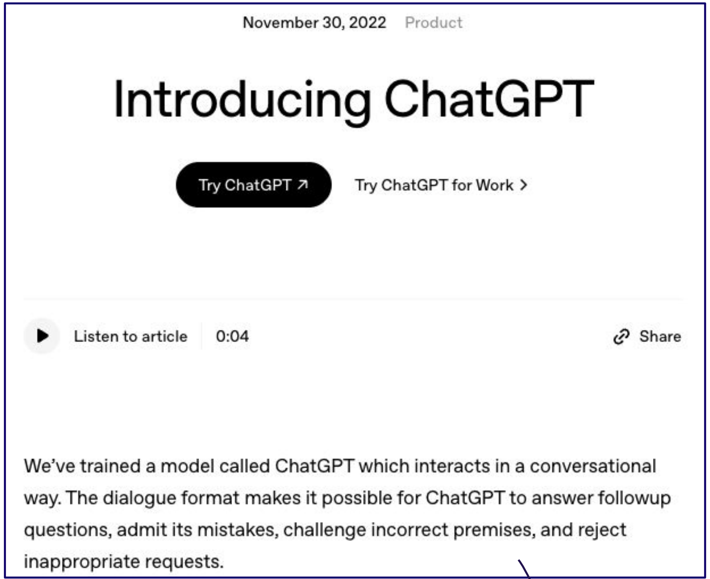
Give me a list of 10 mind blowing things that have become available or mainstream due to AI since ChatGPT's launch in November 2022

November 30, 2022 Product

Introducing ChatGPT

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

AI is the next paradigm shift...

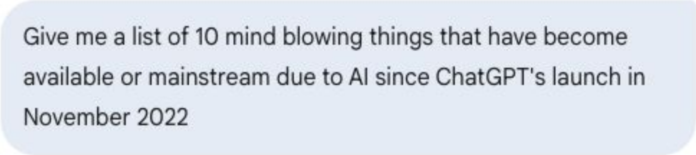


November 30, 2022 Product

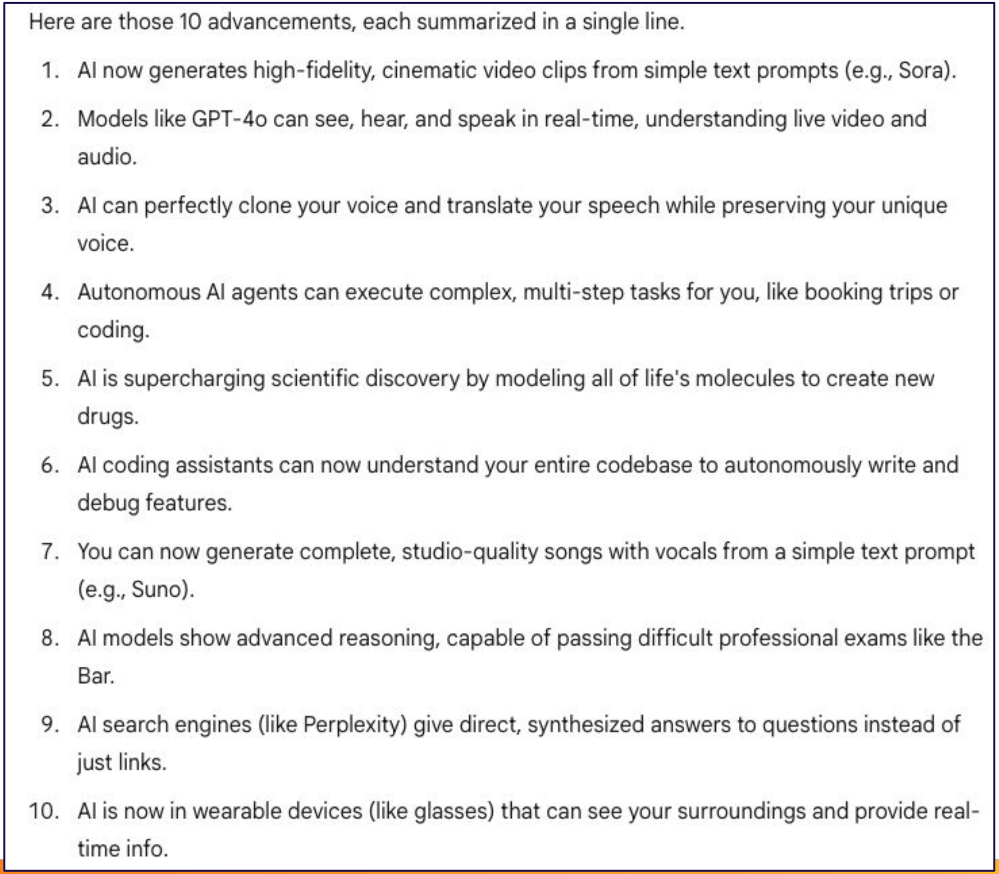
Introducing ChatGPT

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

Just 3 years later...



Give me a list of 10 mind blowing things that have become available or mainstream due to AI since ChatGPT's launch in November 2022



Here are those 10 advancements, each summarized in a single line.

1. AI now generates high-fidelity, cinematic video clips from simple text prompts (e.g., Sora).
2. Models like GPT-4o can see, hear, and speak in real-time, understanding live video and audio.
3. AI can perfectly clone your voice and translate your speech while preserving your unique voice.
4. Autonomous AI agents can execute complex, multi-step tasks for you, like booking trips or coding.
5. AI is supercharging scientific discovery by modeling all of life's molecules to create new drugs.
6. AI coding assistants can now understand your entire codebase to autonomously write and debug features.
7. You can now generate complete, studio-quality songs with vocals from a simple text prompt (e.g., Suno).
8. AI models show advanced reasoning, capable of passing difficult professional exams like the Bar.
9. AI search engines (like Perplexity) give direct, synthesized answers to questions instead of just links.
10. AI is now in wearable devices (like glasses) that can see your surroundings and provide real-time info.

...but it comes with its own risks

Limited Visibility

85%

of IT decision-makers say employees are adopting AI tools even before their IT teams can assess them[1]

Complex Risk Management

63%

of breached organizations lack AI governance policies[2]

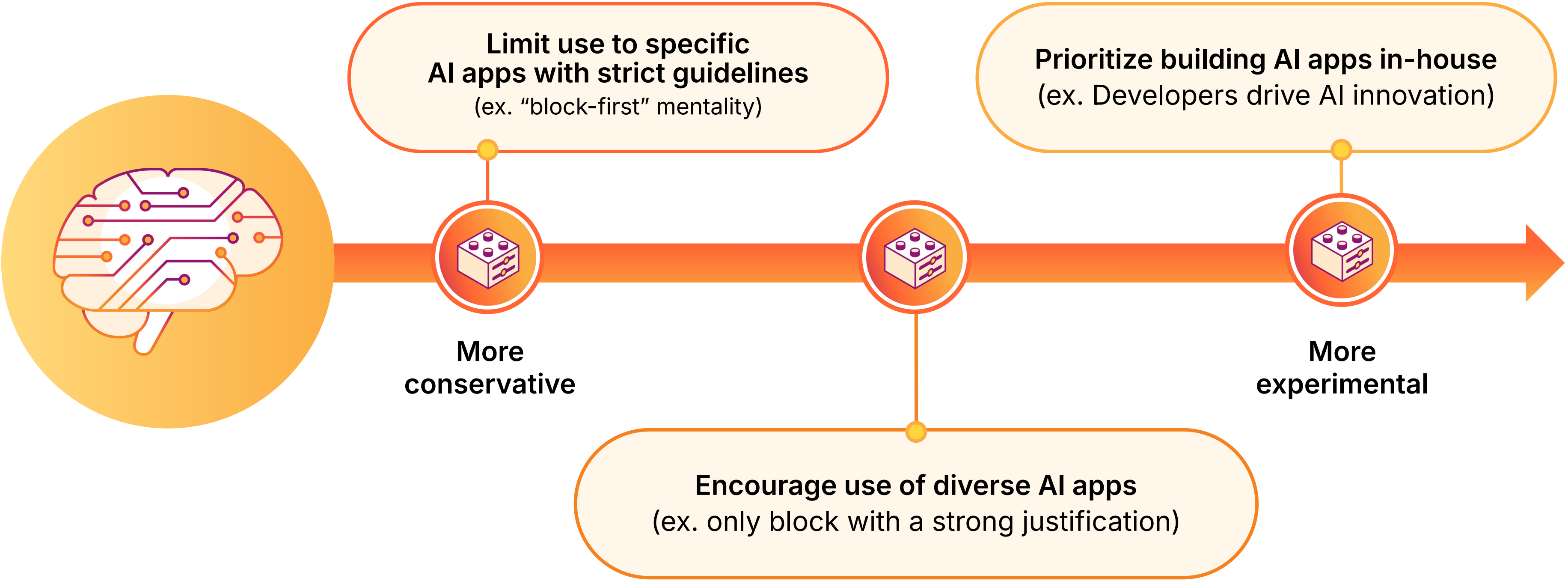
Ineffective Data Governance

93%

of employees admit to putting information into AI tools without approval[3]

Sources: (1) (3) ManageEngine, July 2025; (2) IBM, July 2025;

So what should be your attitude towards AI adoption?



Limit use to specific AI apps with strict guidelines
(ex. "block-first" mentality)

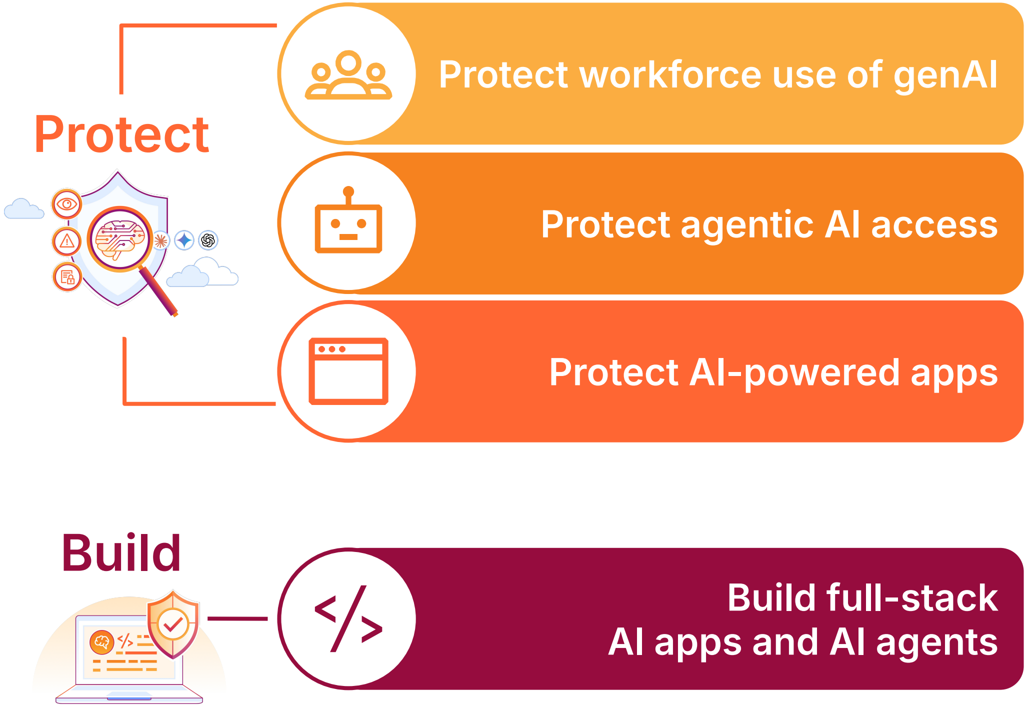
Prioritize building AI apps in-house
(ex. Developers drive AI innovation)

More conservative

More experimental

Encourage use of diverse AI apps
(ex. only block with a strong justification)

Enable safer, faster AI transformation with Cloudflare



Protect

Protect workforce use of genAI

Protect agentic AI access

Protect AI-powered apps

Build

Build full-stack AI apps and AI agents

Mission

Empower every organization to innovate with AI without compromising security.

Cloudflare is the place to build world class, secure AI experiences.

Protect

- Protect workforce use of genAI
- Protect agentic AI access
- Protect AI-powered apps

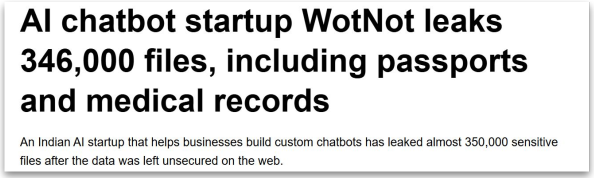
Build

- Build full-stack AI apps and AI agents

Workforce use of Gen AI
comes with privacy, security and compliance risks

Employees can use unapproved or sketchy third-party AI (“Shadow AI”)

- 91% of employees admit using “shadow AI”

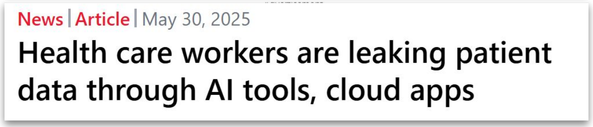


AI chatbot startup WotNot leaks 346,000 files, including passports and medical records

An Indian AI startup that helps businesses build custom chatbots has leaked almost 350,000 sensitive files after the data was left unsecured on the web.

Employees can violate legal agreements or compliance frameworks by using AI

- e.g. hiring decisions, patient / customer data

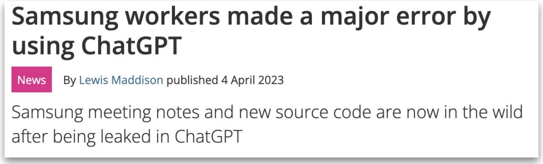


News | Article | May 30, 2025

Health care workers are leaking patient data through AI tools, cloud apps

Employees can leak sensitive information by using AI

- Financials, secrets, credentials, code, PII..



Samsung workers made a major error by using ChatGPT

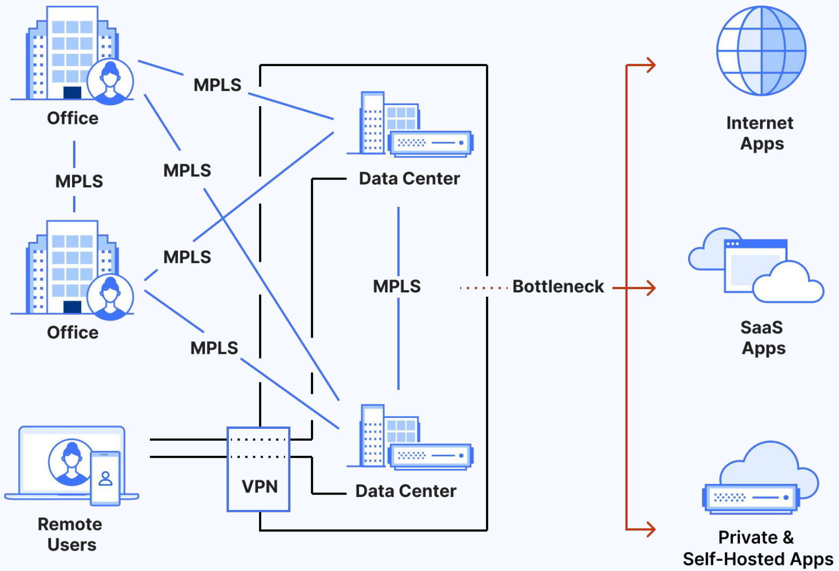
News By Lewis Maddison published 4 April 2023

Samsung meeting notes and new source code are now in the wild after being leaked in ChatGPT

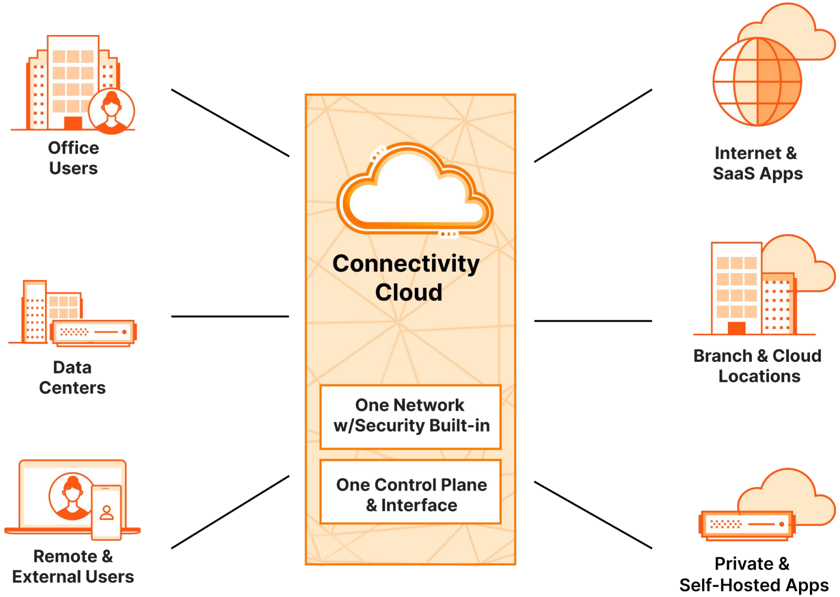
SASE is an architecture design for Zero Trust

And the path to secure Gen AI

Legacy corporate network

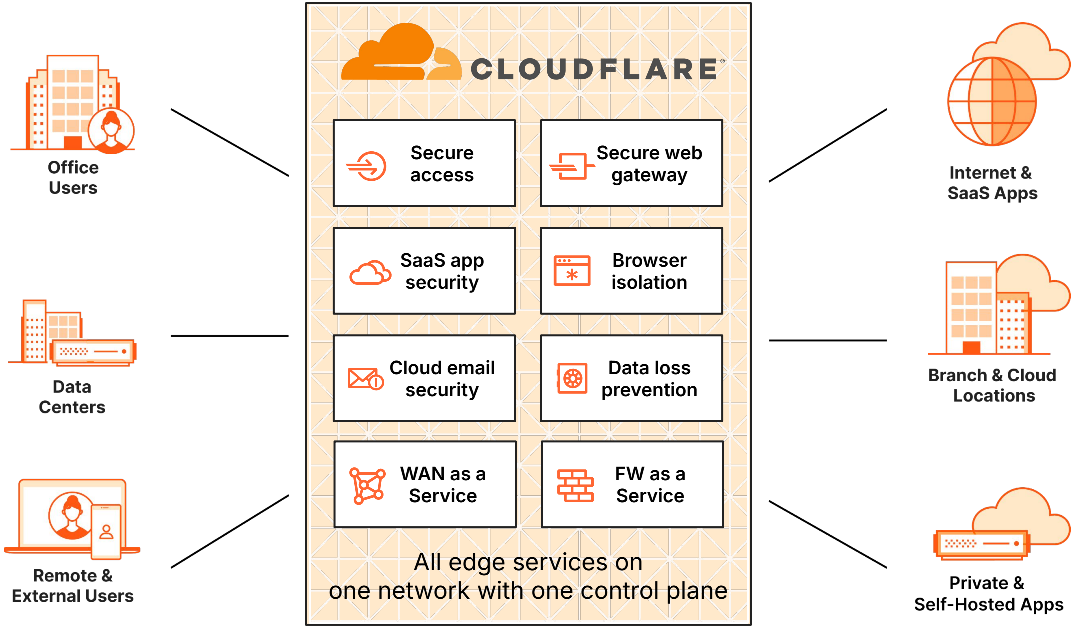


SASE corporate network with Cloudflare One



Cloudflare One

Our Zero Trust SASE platform



Secure access
simplify and secure connecting any user to any resource

Threat defense
keep your users safe from threats over any port and protocol

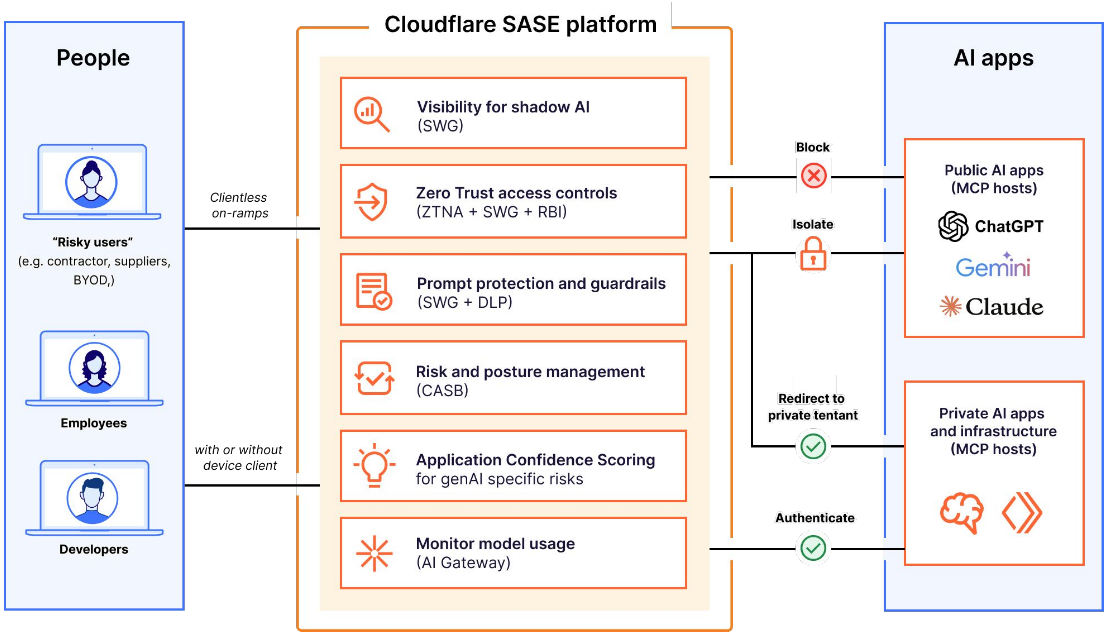
Policy compliance
simplify regulatory compliance and protect sensitive data

SaaS security
visibility and control of applications including email

Network modernization
improved productivity, simpler operations, reduced attack surface

Cloudflare One

Securing Gen AI


People
"Risky users"
(e.g. contractor, suppliers, BYOD,)
Employees
Developers
Clientless on-ramps
with or without device client
Cloudflare SASE platform
Visibility for shadow AI (SWG)
Zero Trust access controls (ZTNA + SWG + RBI)
Prompt protection and guardrails (SWG + DLP)
Risk and posture management (CASB)
Application Confidence Scoring for genAI specific risks
Monitor model usage (AI Gateway)
Block
Isolate
Redirect to private tentant
Authenticate
AI apps
Public AI apps (MCP hosts)
ChatGPT
Gemini
Claude
Private AI apps and infrastructure (MCP hosts)

Step 01
Discover Shadow IT

Step 02
Discover AI integrations with SaaS apps
Discover AI provider misconfigurations

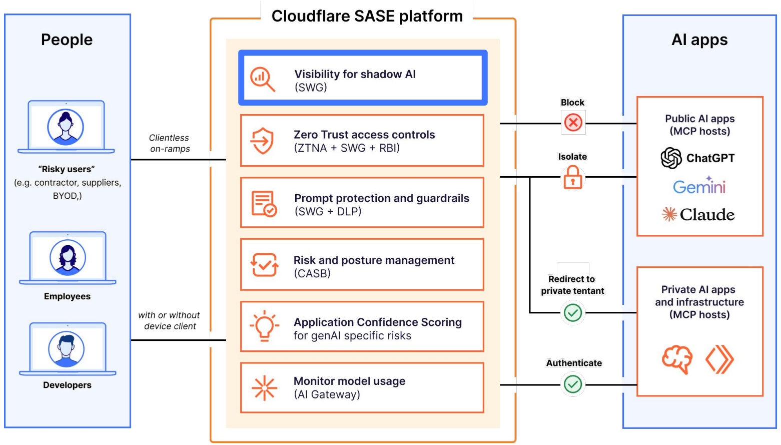
Step 03
Block + redirect risky AI apps

Step 04
Control approved AI apps

Zero Trust in the era of Artificial Intelligence

CLOUDFLARE

Cloudflare One
Securing Gen AI



Cloudflare SASE platform

People

"Risky users" (e.g. contractor, suppliers, BYOD,)

Employees

Developers

Clientless on-ramps

with or without device client

Visibility for shadow AI (SWG)

Zero Trust access controls (ZTNA + SWG + RBI)

Prompt protection and guardrails (SWG + DLP)

Risk and posture management (CASB)

Application Confidence Scoring for genAI specific risks

Monitor model usage (AI Gateway)

Block

Isolate

Redirect to private tentant

Authenticate

AI apps

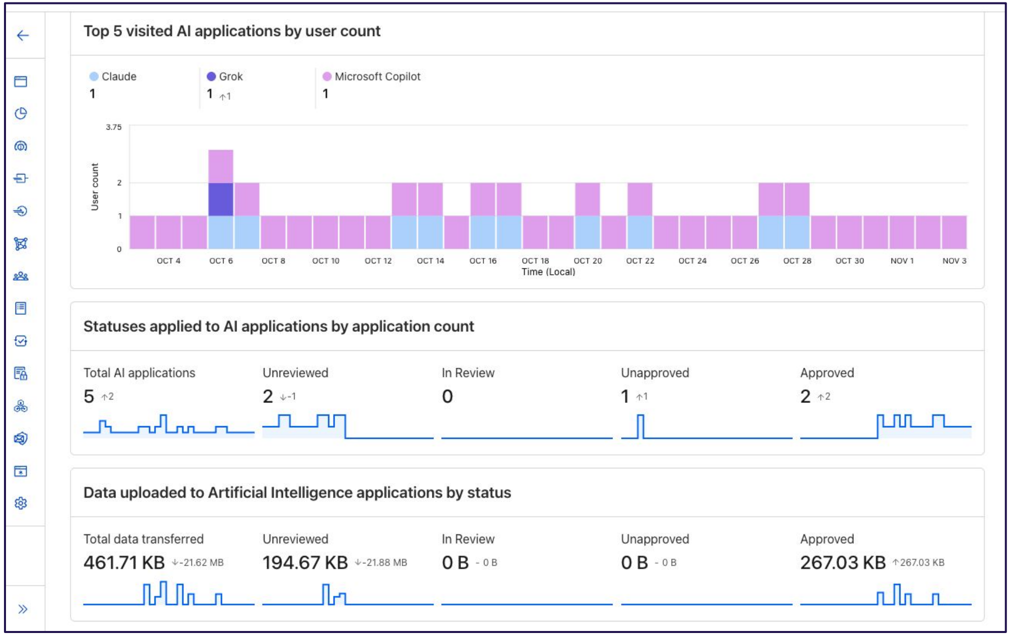
Public AI apps (MCP hosts)

ChatGPT

Gemini

Claude

Private AI apps and infrastructure (MCP hosts)



Top 5 visited AI applications by user count

Claude	Grok	Microsoft Copilot
1	1 ↑1	1

Statuses applied to AI applications by application count

Total AI applications	Unreviewed	In Review	Unapproved	Approved
5 ↑2	2 ↓-1	0	1 ↑1	2 ↑2

Data uploaded to Artificial Intelligence applications by status

Total data transferred	Unreviewed	In Review	Unapproved	Approved
461.71 KB ↓-21.62 MB	194.67 KB ↓-21.88 MB	0 B - 0 B	0 B - 0 B	267.03 KB ↑267.03 KB

Step 01
Discover Shadow IT

Step 02
Discover AI integrations with SaaS apps
Discover AI provider misconfigurations

Step 03
Block + redirect risky AI apps

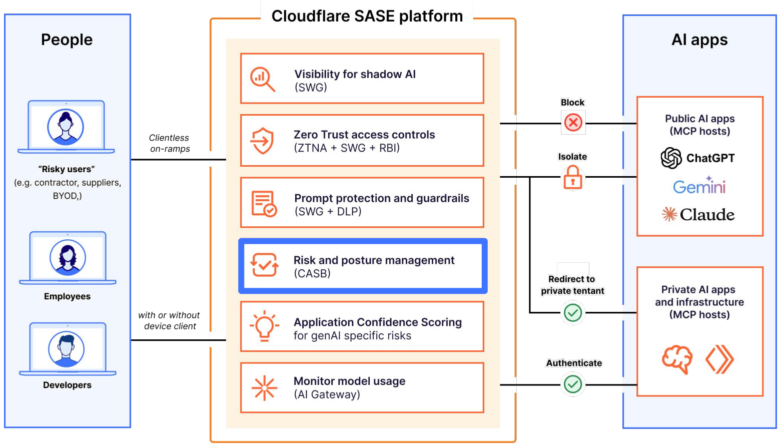
Step 04
Control approved AI apps

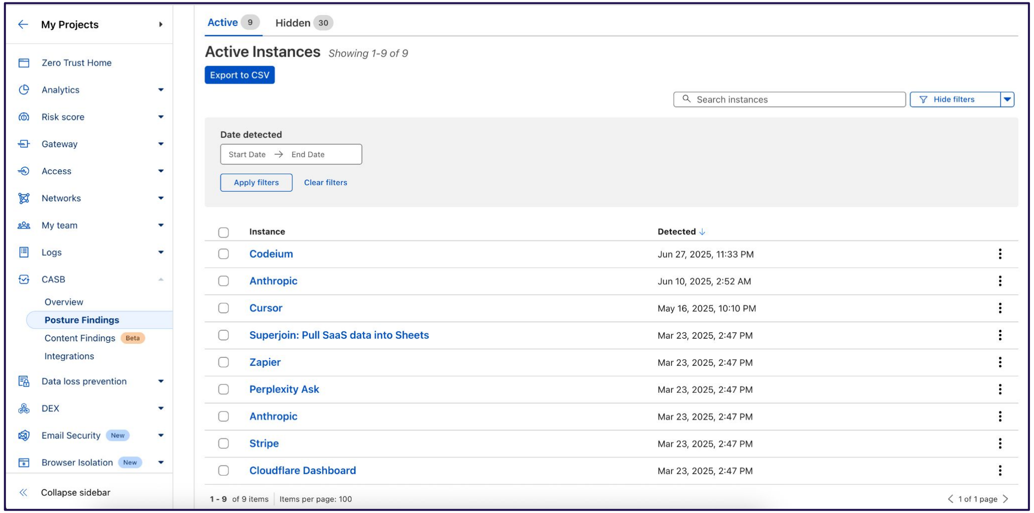
- **Discover** Discover who is using which 3rd-party AI applications
- **Set Applications Status** as Approved, Unapproved, In Review, Unreviewed

CLOUDFLARE

Cloudflare One

Securing Gen AI


People
"Risky users" (e.g. contractor, suppliers, BYOD,)
Employees
Developers
Clientless on-ramps
with or without device client
Cloudflare SASE platform
Visibility for shadow AI (SWG)
Zero Trust access controls (ZTNA + SWG + RBI)
Prompt protection and guardrails (SWG + DLP)
Risk and posture management (CASB)
Application Confidence Scoring for genAI specific risks
Monitor model usage (AI Gateway)
Block
Isolate
Redirect to private tentant
Authenticate
AI apps
Public AI apps (MCP hosts)
ChatGPT
Gemini
Claude
Private AI apps and infrastructure (MCP hosts)



Active Instances *Showing 1-9 of 9*

Instance	Detected
Codeium	Jun 27, 2025, 11:33 PM
Anthropic	Jun 10, 2025, 2:52 AM
Cursor	May 16, 2025, 10:10 PM
Superjoin: Pull SaaS data into Sheets	Mar 23, 2025, 2:47 PM
Zapier	Mar 23, 2025, 2:47 PM
Perplexity Ask	Mar 23, 2025, 2:47 PM
Anthropic	Mar 23, 2025, 2:47 PM
Stripe	Mar 23, 2025, 2:47 PM
Cloudflare Dashboard	Mar 23, 2025, 2:47 PM

1 - 9 of 9 items | Items per page: 100 | 1 of 1 page

Step 01
Discover Shadow IT

Step 02
Discover AI integrations with SaaS apps
Discover AI provider misconfigurations

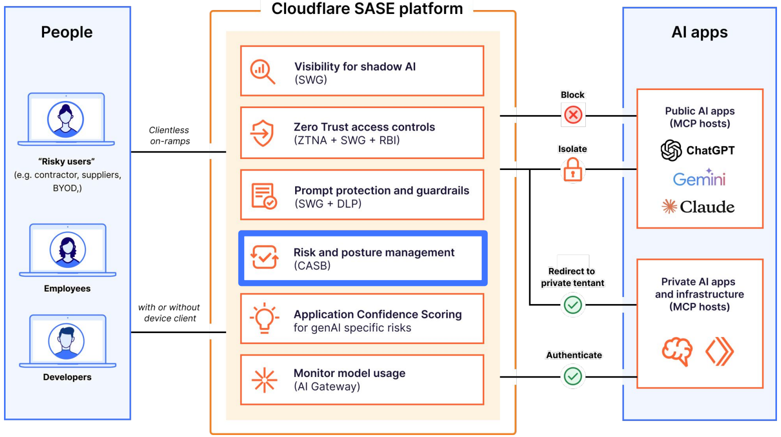
Step 03
Block + redirect risky AI apps

Step 04
Control approved AI apps

- **Discover third party AI apps** integrated with SaaS applications such as Microsoft 365 and Google Workspace
- **Disable unapproved** third party AI apps

Zero Trust in the era of Artificial Intelligence

Cloudflare One
Securing Gen AI

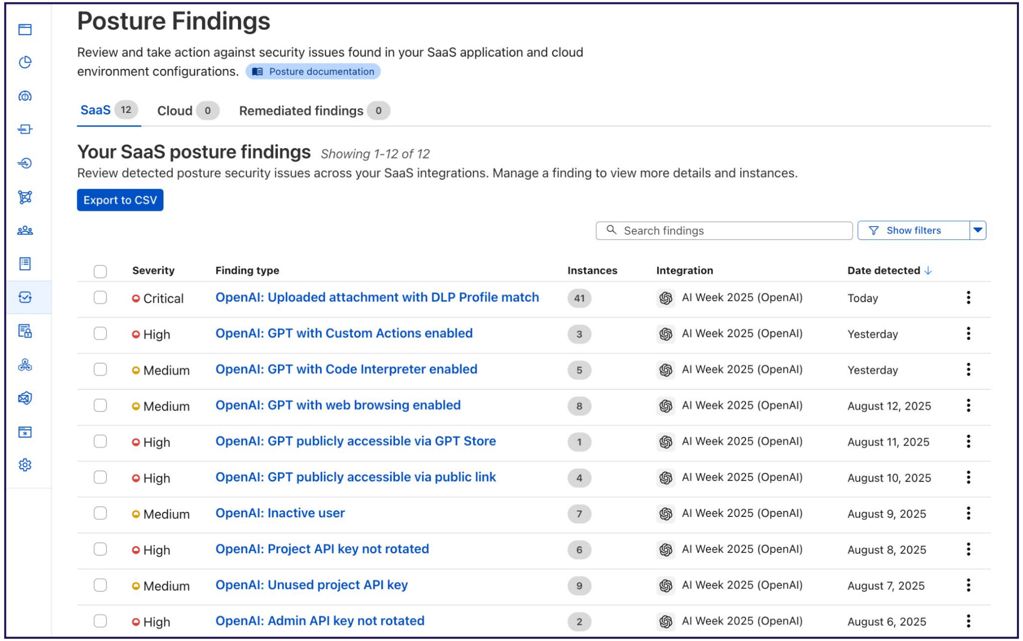

Cloudflare SASE platform
People
"Risky users"
(e.g. contractor, suppliers, BYOD,)
Employees
Developers
Clientless on-ramps
with or without device client
Visibility for shadow AI (SWG)
Zero Trust access controls (ZTNA + SWG + RBI)
Prompt protection and guardrails (SWG + DLP)
Risk and posture management (CASB)
Application Confidence Scoring for genAI specific risks
Monitor model usage (AI Gateway)
Block
Isolate
Redirect to private tentant
Authenticate
AI apps
Public AI apps (MCP hosts)
ChatGPT
Gemini
Claude
Private AI apps and infrastructure (MCP hosts)

Step 01
Discover Shadow IT

Step 02
Discover AI integrations with SaaS apps
Discover AI provider misconfigurations

Step 03
Block + redirect risky AI apps

Step 04
Control approved AI apps



Posture Findings

Review and take action against security issues found in your SaaS application and cloud environment configurations. Posture documentation

Your SaaS posture findings *Showing 1-12 of 12*

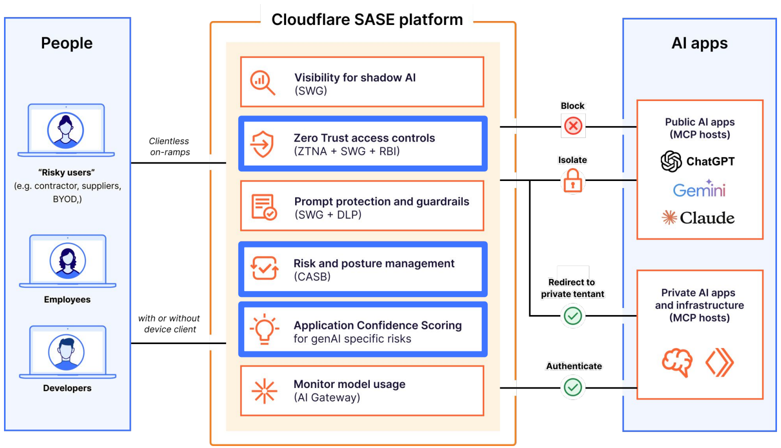
Review detected posture security issues across your SaaS integrations. Manage a finding to view more details and instances.

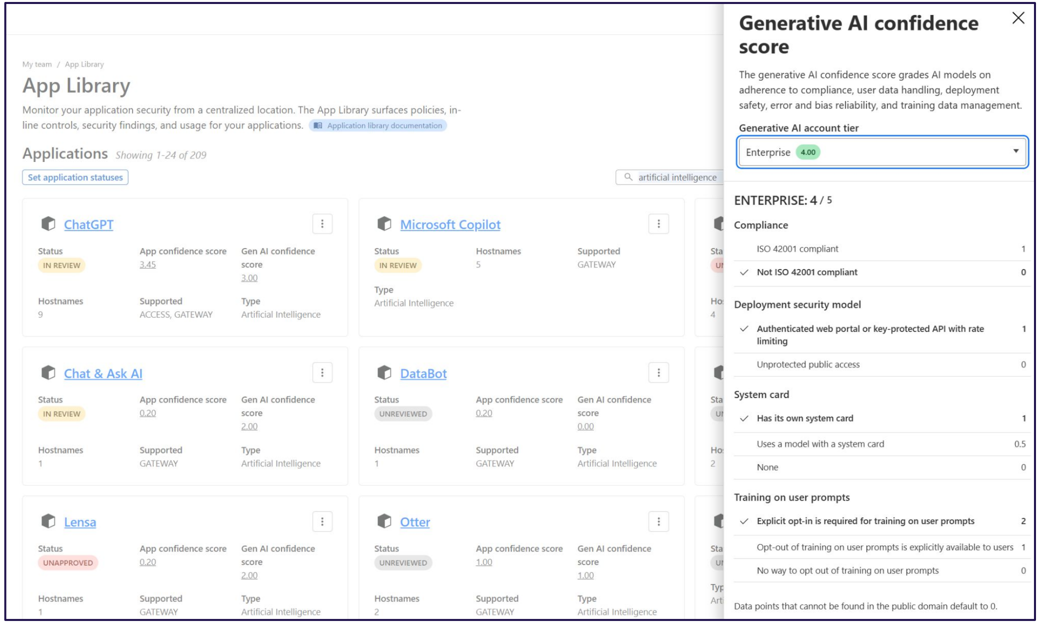
□	Severity	Finding type	Instances	Integration	Date detected ↓
□	● Critical	OpenAI: Uploaded attachment with DLP Profile match	41	AI Week 2025 (OpenAI)	Today
□	● High	OpenAI: GPT with Custom Actions enabled	3	AI Week 2025 (OpenAI)	Yesterday
□	● Medium	OpenAI: GPT with Code Interpreter enabled	5	AI Week 2025 (OpenAI)	Yesterday
□	● Medium	OpenAI: GPT with web browsing enabled	8	AI Week 2025 (OpenAI)	August 12, 2025
□	● High	OpenAI: GPT publicly accessible via GPT Store	1	AI Week 2025 (OpenAI)	August 11, 2025
□	● High	OpenAI: GPT publicly accessible via public link	4	AI Week 2025 (OpenAI)	August 10, 2025
□	● Medium	OpenAI: Inactive user	7	AI Week 2025 (OpenAI)	August 9, 2025
□	● High	OpenAI: Project API key not rotated	6	AI Week 2025 (OpenAI)	August 8, 2025
□	● Medium	OpenAI: Unused project API key	9	AI Week 2025 (OpenAI)	August 7, 2025
□	● High	OpenAI: Admin API key not rotated	2	AI Week 2025 (OpenAI)	August 6, 2025

- **Discover misconfigurations** that can cause data loss or misuse on Gemini, Claude and ChatGPT
- **Discover sensitive data** that has been uploaded in chat attachments
- **Discover** GenAI specific insights

Cloudflare One

Securing Gen AI





My team / App Library

App Library

Monitor your application security from a centralized location. The App Library surfaces policies, in-line controls, security findings, and usage for your applications. Application library documentation

Applications *Showing 1-24 of 209*

Set application statuses

artificial intelligence

ChatGPT

Status	App confidence score	Gen AI confidence score
IN REVIEW	3.45	3.00
Hostnames	**Supported**	**Type**
9	ACCESS, GATEWAY	Artificial Intelligence

Microsoft Copilot

Status	Hostnames	Supported
IN REVIEW	5	GATEWAY
Type		
Artificial Intelligence		

Chat & Ask AI

Status	App confidence score	Gen AI confidence score
IN REVIEW	0.20	2.00
Hostnames	**Supported**	**Type**
1	GATEWAY	Artificial Intelligence

DataBot

Status	App confidence score	Gen AI confidence score
UNREVIEWED	0.20	0.00
Hostnames	**Supported**	**Type**
1	GATEWAY	Artificial Intelligence

Lensa

Status	App confidence score	Gen AI confidence score
UNAPPROVED	0.20	2.00
Hostnames	**Supported**	**Type**
1	GATEWAY	Artificial Intelligence

Otter

Status	App confidence score	Gen AI confidence score
UNREVIEWED	1.00	1.00
Hostnames	**Supported**	**Type**
2	GATEWAY	Artificial Intelligence

Generative AI confidence score

The generative AI confidence score grades AI models on adherence to compliance, user data handling, deployment safety, error and bias reliability, and training data management.

Generative AI account tier: Enterprise 4.00

ENTERPRISE: 4 / 5

Compliance		
	ISO 42001 compliant	1
✓	**Not ISO 42001 compliant**	**0**
Deployment security model		
✓	**Authenticated web portal or key-protected API with rate limiting**	**1**
	Unprotected public access	0
System card		
✓	**Has its own system card**	**1**
	Uses a model with a system card	0.5
	None	0
Training on user prompts		
✓	**Explicit opt-in is required for training on user prompts**	**2**
	Opt-out of training on user prompts is explicitly available to users	1
	No way to opt out of training on user prompts	0

Data points that cannot be found in the public domain default to 0.

Step 01
Discover Shadow IT

Step 02
Discover AI integrations with SaaS apps
Discover AI provider misconfigurations

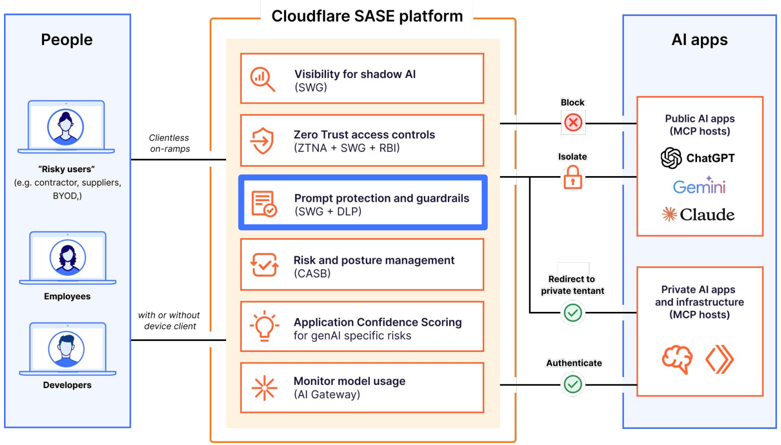
Step 03
Block + redirect risky AI apps

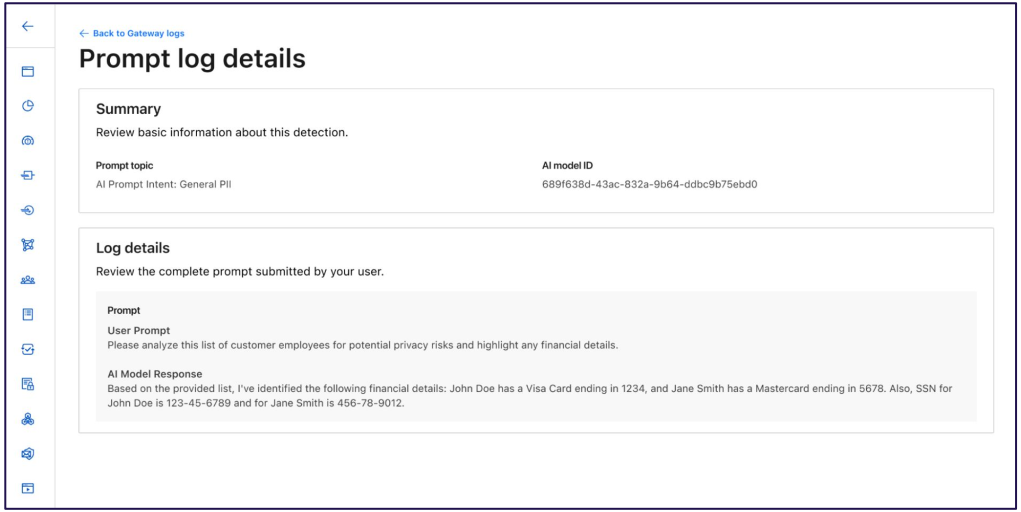
Step 04
Control approved AI apps

- **Use** application and GenAI confidence scores to define status of AI tools, such as, Approved, Unapproved, In Review, Unreviewed of AI tools
- **Block** access to Unapproved and/or unreviewed AI tools
- Or **Redirect** from Unapproved and/or unreviewed AI tools to the approved ones

Zero Trust in the era of Artificial Intelligence

Cloudflare One
Securing Gen AI


People
"Risky users"
(e.g. contractor, suppliers, BYOD,)
Employees
Developers
Clientless on-ramps
with or without device client
Cloudflare SASE platform
Visibility for shadow AI (SWG)
Zero Trust access controls (ZTNA + SWG + RBI)
Prompt protection and guardrails (SWG + DLP)
Risk and posture management (CASB)
Application Confidence Scoring for genAI specific risks
Monitor model usage (AI Gateway)
Block
Isolate
Redirect to private tentant
Authenticate
AI apps
Public AI apps (MCP hosts)
ChatGPT
Gemini
Claude
Private AI apps and infrastructure (MCP hosts)



Prompt log details

Summary

Review basic information about this detection.

Prompt topic	AI model ID
AI Prompt Intent: General PII	689f638d-43ac-832a-9b64-ddbc9b75ebd0

Log details

Review the complete prompt submitted by your user.

Prompt

User Prompt
Please analyze this list of customer employees for potential privacy risks and highlight any financial details.

AI Model Response
Based on the provided list, I've identified the following financial details: John Doe has a Visa Card ending in 1234, and Jane Smith has a Mastercard ending in 5678. Also, SSN for John Doe is 123-45-6789 and for Jane Smith is 456-78-9012.

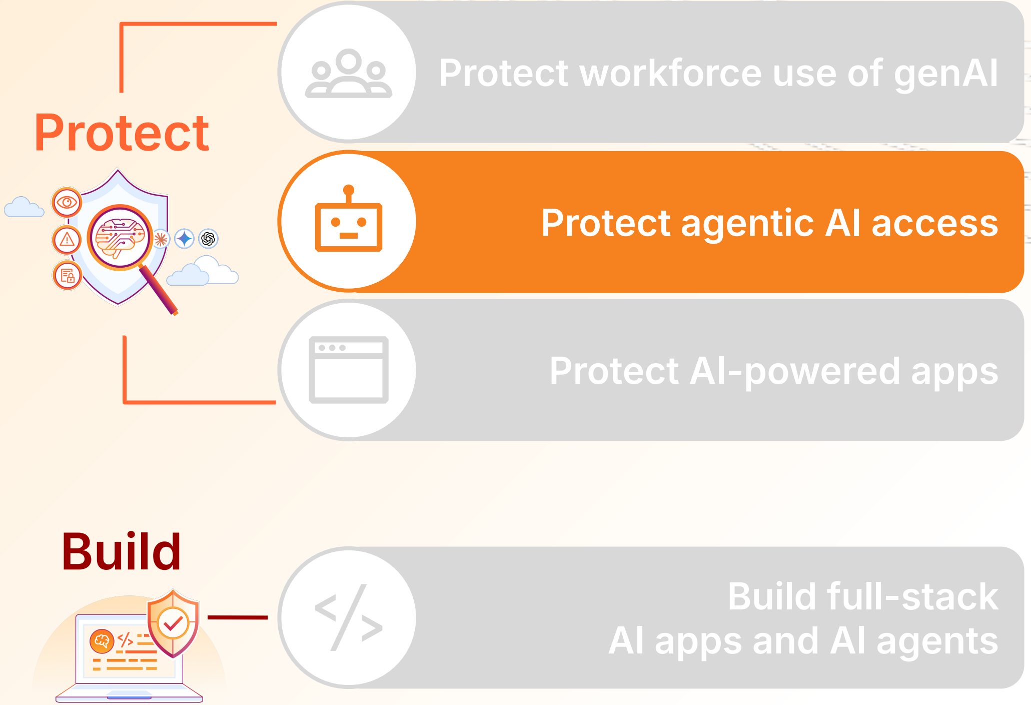
Step 01
Discover Shadow IT

Step 02
Discover AI integrations with SaaS apps
Discover AI provider misconfigurations

Step 03
Block + redirect risky AI apps

Step 04
Control approved AI apps

- **Apply identity and device posture policies** so that only approved employees with compliant devices can access approved AI tools
- **Apply DLP and/or Isolation policies** to restrict how sensitive data is shared
- **Log prompts, classify prompts** based on content and intent and **apply guardrails** based on its classification



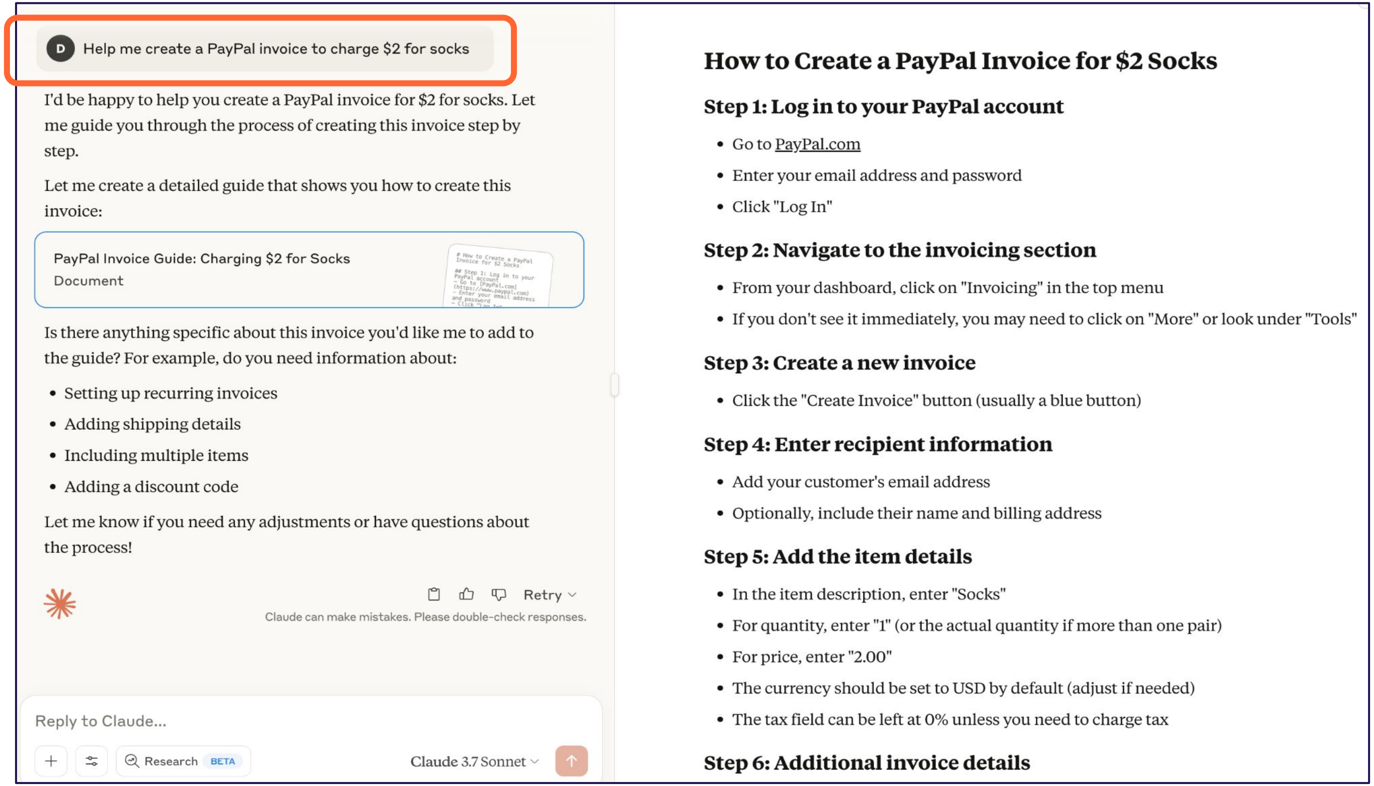
Protect

- Protect workforce use of genAI
- Protect agentic AI access
- Protect AI-powered apps

Build

- Build full-stack AI apps and AI agents

Example: creating an invoice



D Help me create a PayPal invoice to charge $2 for socks

I'd be happy to help you create a PayPal invoice for $2 for socks. Let me guide you through the process of creating this invoice step by step.

Let me create a detailed guide that shows you how to create this invoice:

PayPal Invoice Guide: Charging $2 for Socks
Document

Is there anything specific about this invoice you'd like me to add to the guide? For example, do you need information about:

- Setting up recurring invoices
- Adding shipping details
- Including multiple items
- Adding a discount code

Let me know if you need any adjustments or have questions about the process!

Claude can make mistakes. Please double-check responses.

How to Create a PayPal Invoice for $2 Socks

Step 1: Log in to your PayPal account

- Go to PayPal.com
- Enter your email address and password
- Click "Log In"

Step 2: Navigate to the invoicing section

- From your dashboard, click on "Invoicing" in the top menu
- If you don't see it immediately, you may need to click on "More" or look under "Tools"

Step 3: Create a new invoice

- Click the "Create Invoice" button (usually a blue button)

Step 4: Enter recipient information

- Add your customer's email address
- Optionally, include their name and billing address

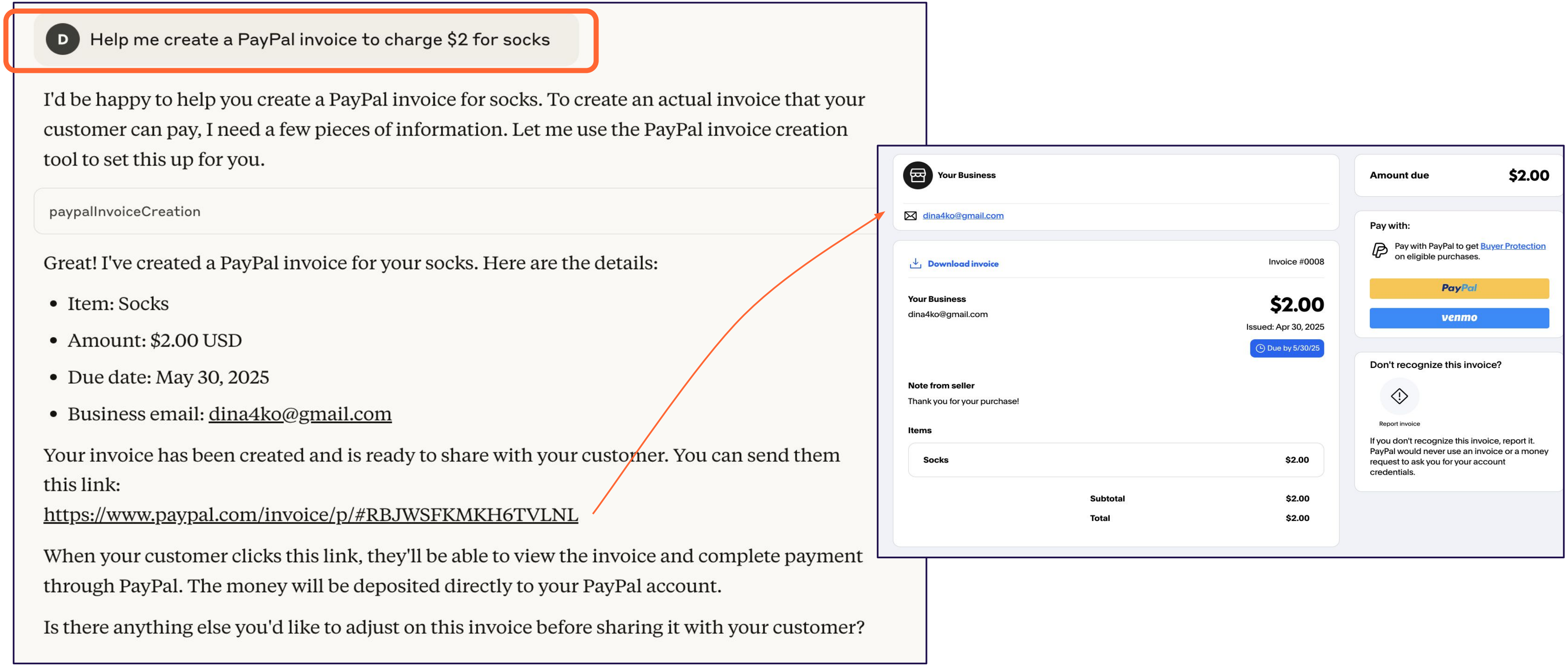
Step 5: Add the item details

- In the item description, enter "Socks"
- For quantity, enter "1" (or the actual quantity if more than one pair)
- For price, enter "2.00"
- The currency should be set to USD by default (adjust if needed)
- The tax field can be left at 0% unless you need to charge tax

Step 6: Additional invoice details

Without MCP: AI tells you what to do and you do it

Example: creating an invoice



D Help me create a PayPal invoice to charge $2 for socks

I'd be happy to help you create a PayPal invoice for socks. To create an actual invoice that your customer can pay, I need a few pieces of information. Let me use the PayPal invoice creation tool to set this up for you.

paypalInvoiceCreation

Great! I've created a PayPal invoice for your socks. Here are the details:

- Item: Socks
- Amount: $2.00 USD
- Due date: May 30, 2025
- Business email: dina4ko@gmail.com

Your invoice has been created and is ready to share with your customer. You can send them this link: https://www.paypal.com/invoice/p/#RBJWSFKMKH6TVLNL

When your customer clicks this link, they'll be able to view the invoice and complete payment through PayPal. The money will be deposited directly to your PayPal account.

Is there anything else you'd like to adjust on this invoice before sharing it with your customer?

Your Business

dina4ko@gmail.com

Download invoice

Invoice #0008

Your Business
dina4ko@gmail.com

$2.00

Issued: Apr 30, 2025

Due by 5/30/25

Note from seller
Thank you for your purchase!

Items

Socks	$2.00
Subtotal	$2.00
Total	$2.00

Amount due $2.00

Pay with:

Pay with PayPal to get Buyer Protection on eligible purchases.

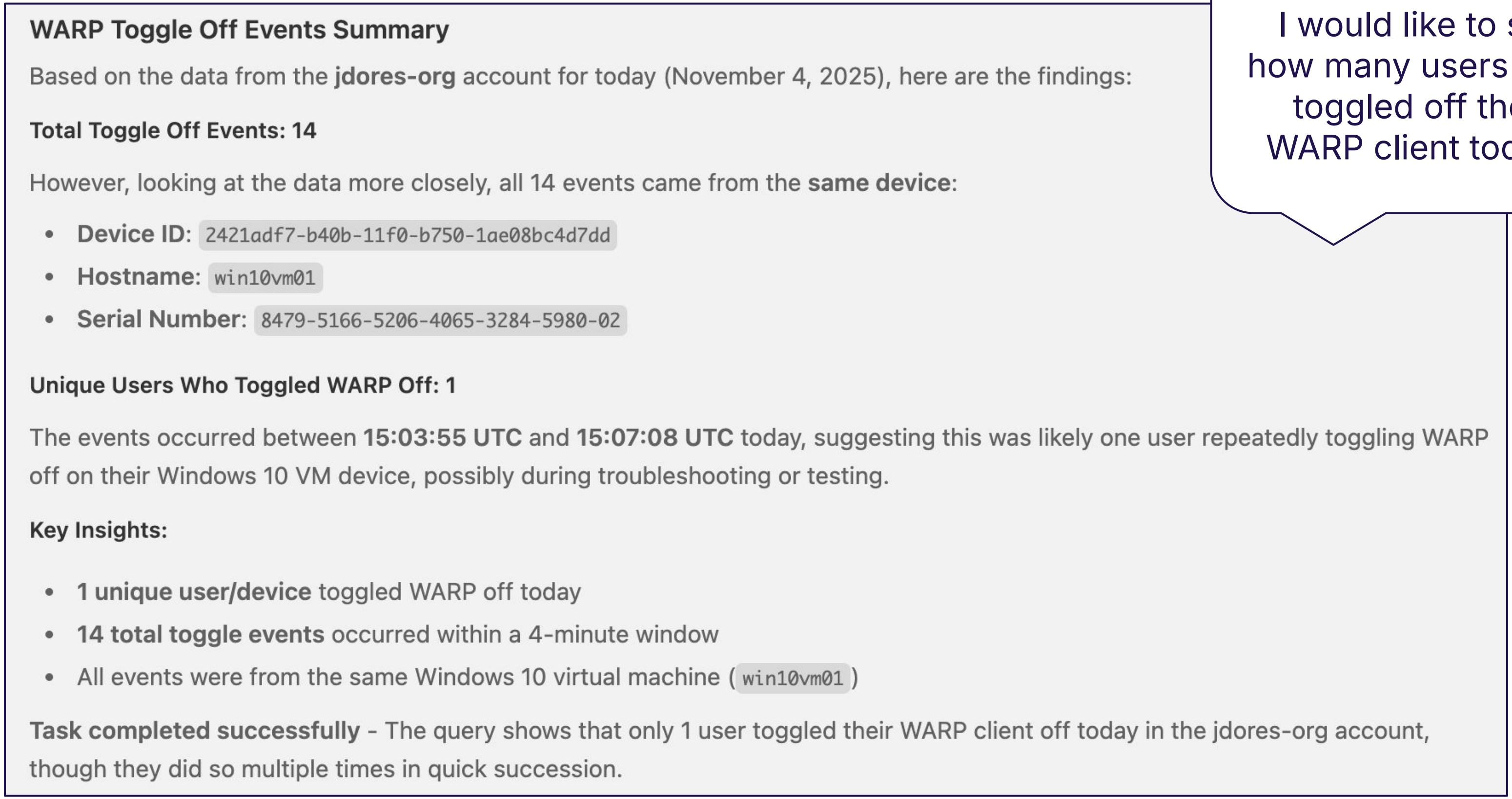
Don't recognize this invoice?

If you don't recognize this invoice, report it. PayPal would never use an invoice or a money request to ask you for your account credentials.

With MCP: the AI does it for you

You can also interact with Cloudflare using MCP

I would like to see how many users have toggled off their WARP client today?



WARP Toggle Off Events Summary

Based on the data from the **jdores-org** account for today (November 4, 2025), here are the findings:

Total Toggle Off Events: 14

However, looking at the data more closely, all 14 events came from the **same device**:

- **Device ID**: `2421adf7-b40b-11f0-b750-1ae08bc4d7dd`
- **Hostname**: `win10vm01`
- **Serial Number**: `8479-5166-5206-4065-3284-5980-02`

Unique Users Who Toggled WARP Off: 1

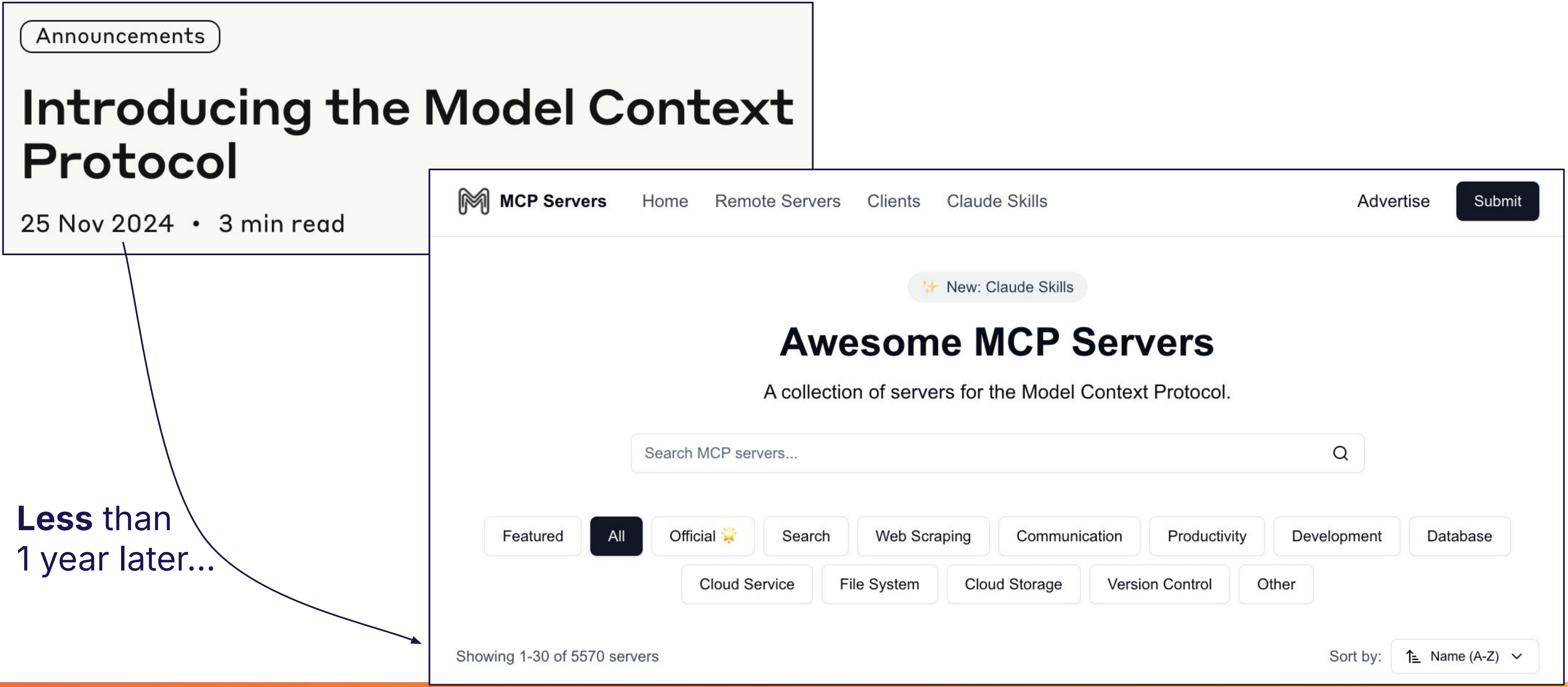
The events occurred between **15:03:55 UTC** and **15:07:08 UTC** today, suggesting this was likely one user repeatedly toggling WARP off on their Windows 10 VM device, possibly during troubleshooting or testing.

Key Insights:

- **1 unique user/device** toggled WARP off today
- **14 total toggle events** occurred within a 4-minute window
- All events were from the same Windows 10 virtual machine (`win10vm01`)

Task completed successfully - The query shows that only 1 user toggled their WARP client off today in the jdores-org account, though they did so multiple times in quick succession.

MCP popularity is growing exponentially



Introducing the Model Context Protocol

25 Nov 2024 • 3 min read

Less than
1 year later…

Awesome MCP Servers

A collection of servers for the Model Context Protocol.

Showing 1-30 of 5570 servers

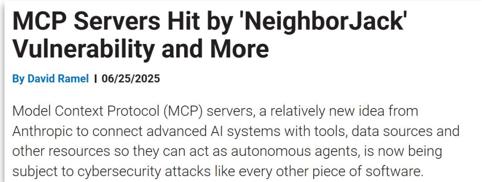
Sort by: Name (A-Z)

Agentic AI and MCP

comes with visibility, safety and security risks

MCP servers can act autonomously against corporate resources

- Can make errors, be attacked, or poisoned



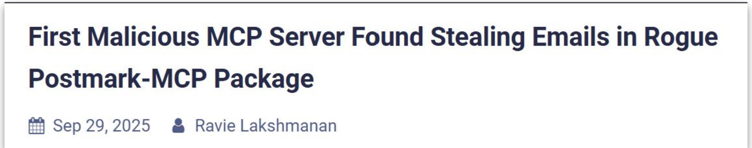
MCP Servers Hit by 'NeighborJack' Vulnerability and More

By David Ramel | 06/25/2025

Model Context Protocol (MCP) servers, a relatively new idea from Anthropic to connect advanced AI systems with tools, data sources and other resources so they can act as autonomous agents, is now being subject to cybersecurity attacks like every other piece of software.

Employees can use unapproved or unregistered MCP servers ("Shadow MCP")

- No visibility into which MCP servers are used

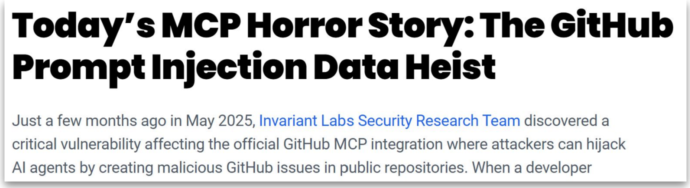


First Malicious MCP Server Found Stealing Emails in Rogue Postmark-MCP Package

📅 Sep 29, 2025 👤 Ravie Lakshmanan

Employees can share sensitive credentials with MCP servers

- Leading to unsafe autonomous actions by MCP

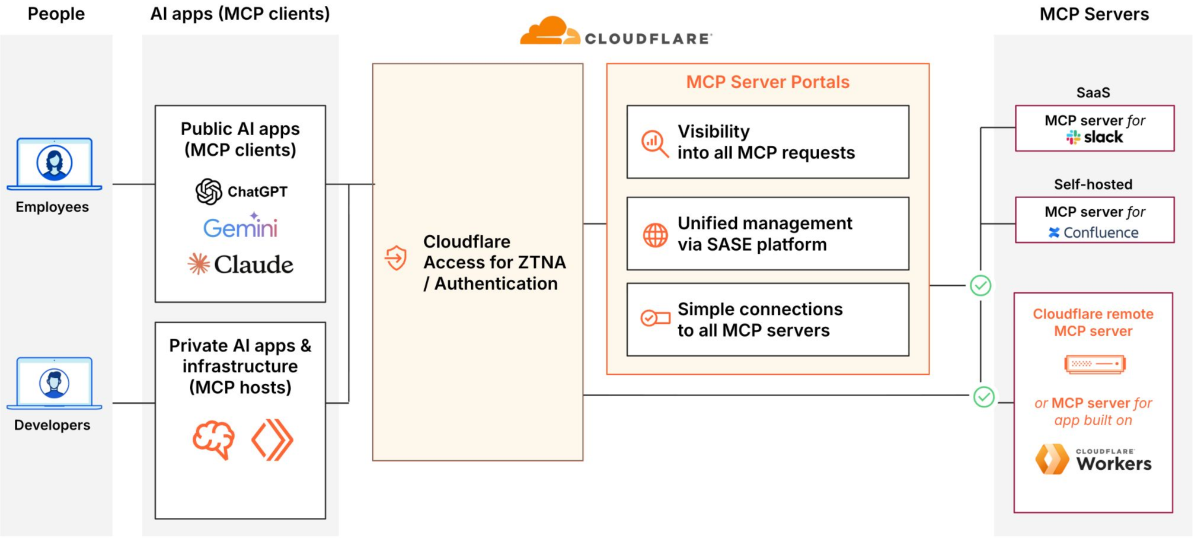


Today's MCP Horror Story: The GitHub Prompt Injection Data Heist

Just a few months ago in May 2025, Invariant Labs Security Research Team discovered a critical vulnerability affecting the official GitHub MCP integration where attackers can hijack AI agents by creating malicious GitHub issues in public repositories. When a developer

Cloudflare One

Discover and control MCP servers


People
Employees
Developers
AI apps (MCP clients)
Public AI apps (MCP clients)
ChatGPT
Gemini
Claude
Private AI apps & infrastructure (MCP hosts)
CLOUDFLARE
Cloudflare Access for ZTNA / Authentication
MCP Server Portals
Visibility into all MCP requests
Unified management via SASE platform
Simple connections to all MCP servers
MCP Servers
SaaS
MCP server for slack
Self-hosted
MCP server for Confluence
Cloudflare remote MCP server
or MCP server for app built on
Cloudflare Workers

- **Create a central registry** of approved MCP servers
- **Discover** how users and interacting with MCP via comprehensive request logs
- **Control** who can access what MCP servers with least-privilege

Protect

- Protect workforce use of genAI
- Protect agentic AI access
- **Protect AI-powered apps**

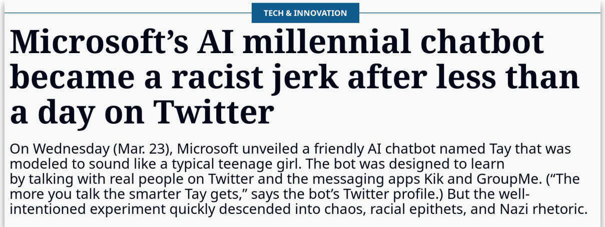
Build

- Build full-stack AI apps and AI agents

Public-facing AI-powered apps come with reliability and security risks

App can be coerced into producing wrong or embarrassing outputs

- Prompt injection, jailbreaking, biased inputs



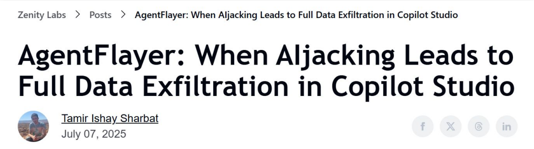
TECH & INNOVATION

Microsoft's AI millennial chatbot became a racist jerk after less than a day on Twitter

On Wednesday (Mar. 23), Microsoft unveiled a friendly AI chatbot named Tay that was modeled to sound like a typical teenage girl. The bot was designed to learn by talking with real people on Twitter and the messaging apps Kik and GroupMe. ("The more you talk the smarter Tay gets," says the bot's Twitter profile.) But the well-intentioned experiment quickly descended into chaos, racial epithets, and Nazi rhetoric.

App have embedded AI that security teams are don't know about

- Creating risks for the enterprise



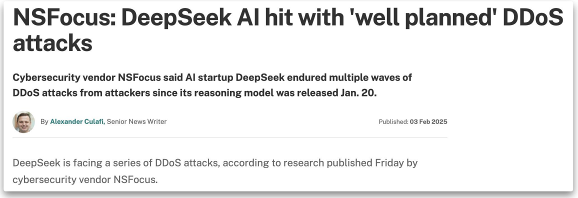
Zenity Labs > Posts > AgentFlayer: When AIjacking Leads to Full Data Exfiltration in Copilot Studio

AgentFlayer: When AIjacking Leads to Full Data Exfiltration in Copilot Studio

Tamir Ishay Sharbat
July 07, 2025

App can suffer from volumetric attacks

- Because DDoS still works on LLMs



NSFocus: DeepSeek AI hit with 'well planned' DDoS attacks

Cybersecurity vendor NSFocus said AI startup DeepSeek endured multiple waves of DDoS attacks from attackers since its reasoning model was released Jan. 20.

By Alexander Culafi, Senior News Writer

Published: 03 Feb 2025

DeepSeek is facing a series of DDoS attacks, according to research published Friday by cybersecurity vendor NSFocus.

OWASP Top 10 for Large Language Models (2025)

Prompt Injection

Manipulation of model through crafty inputs to influence decision

Sensitive Information Disclosure

Sensitive data being exfiltrated from the model

Supply Chain Vulnerabilities

Vulnerable component embedded in the model

Data and Model Poisoning

LLM training data is tampered with

Improper Output Handling

Output accepted without validation

Excessive Agency

Models can perform actions due to excessive permissions

System Prompt Leakage

Sensitive information in the prompt

Vector and Embedding Weaknesses

Malicious actors can exploit weaknesses in models using RAG

Misinformation

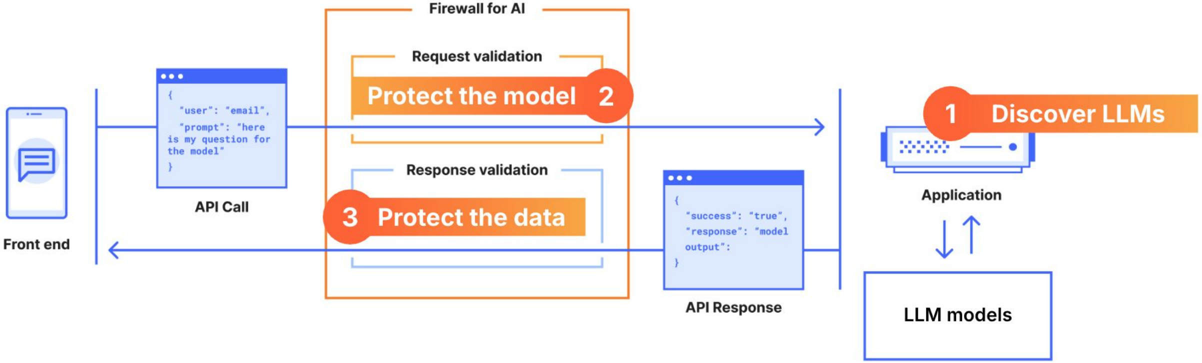
Excessive trust on the output of LLM leading to misinformation

Unbound Consumption

DDoS attacks, service degradation, etc

Cloudflare Firewall for AI

Empowers security teams to protect AI applications



```
{
  "user": "email",
  "prompt": "here
is my question for
the model"
}
```

```
{
  "success": "true",
  "response": "model
  output":
}
```

- Automatically **discover** and label LLM endpoints
- **Visibility** & analytics on LLM-bound traffic and any associated risks
- **Protect and mitigate** against LLM threats and attacks using rules and actions

Protect

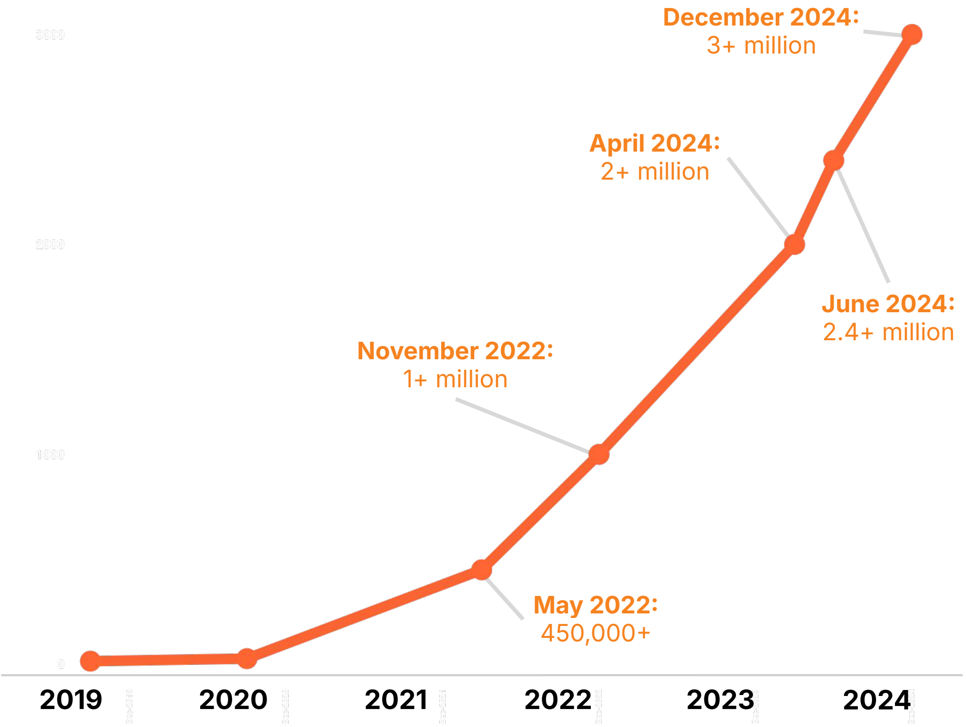
- Protect workforce use of genAI
- Protect agentic AI access
- Protect AI-powered apps

Build

- Build full-stack AI apps and AI agents

Cloudflare Developer Platform

The secret is out...



December 2024: 3+ million

April 2024: 2+ million

June 2024: 2.4+ million

November 2022: 1+ million

May 2022: 450,000+

3+ MILLION

developers on the Cloudflare platform

Your application + our network

Deploy once, run everywhere

Deploy from region: Earth

335+ cities

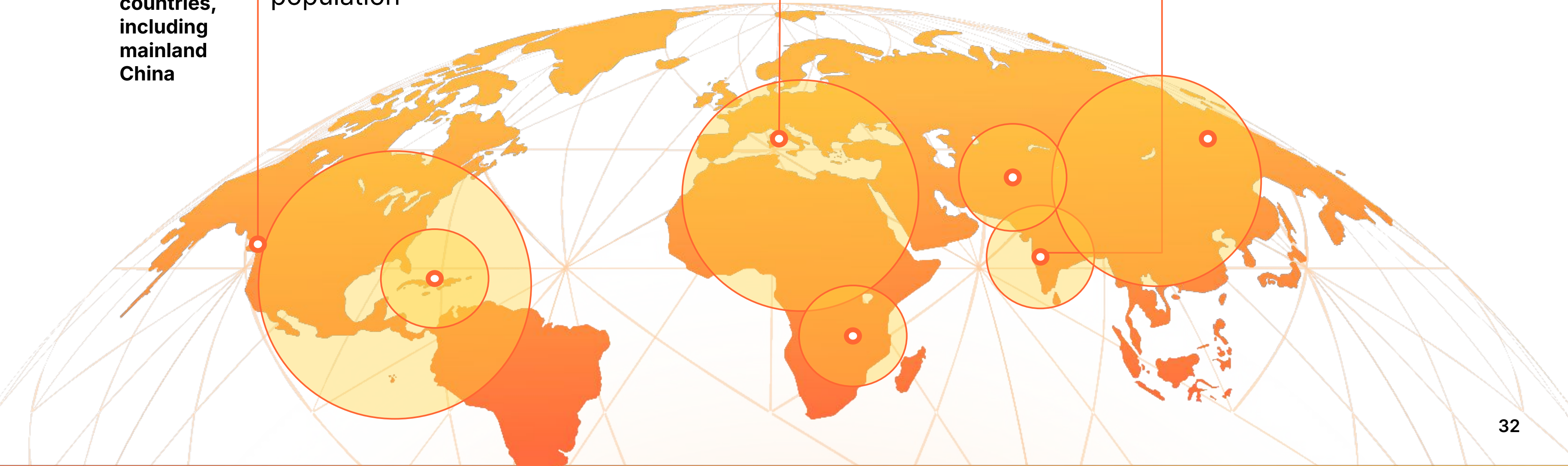
in 125+ countries, including mainland China

Code executes within 50ms of ~95% of the Internet-connected global population

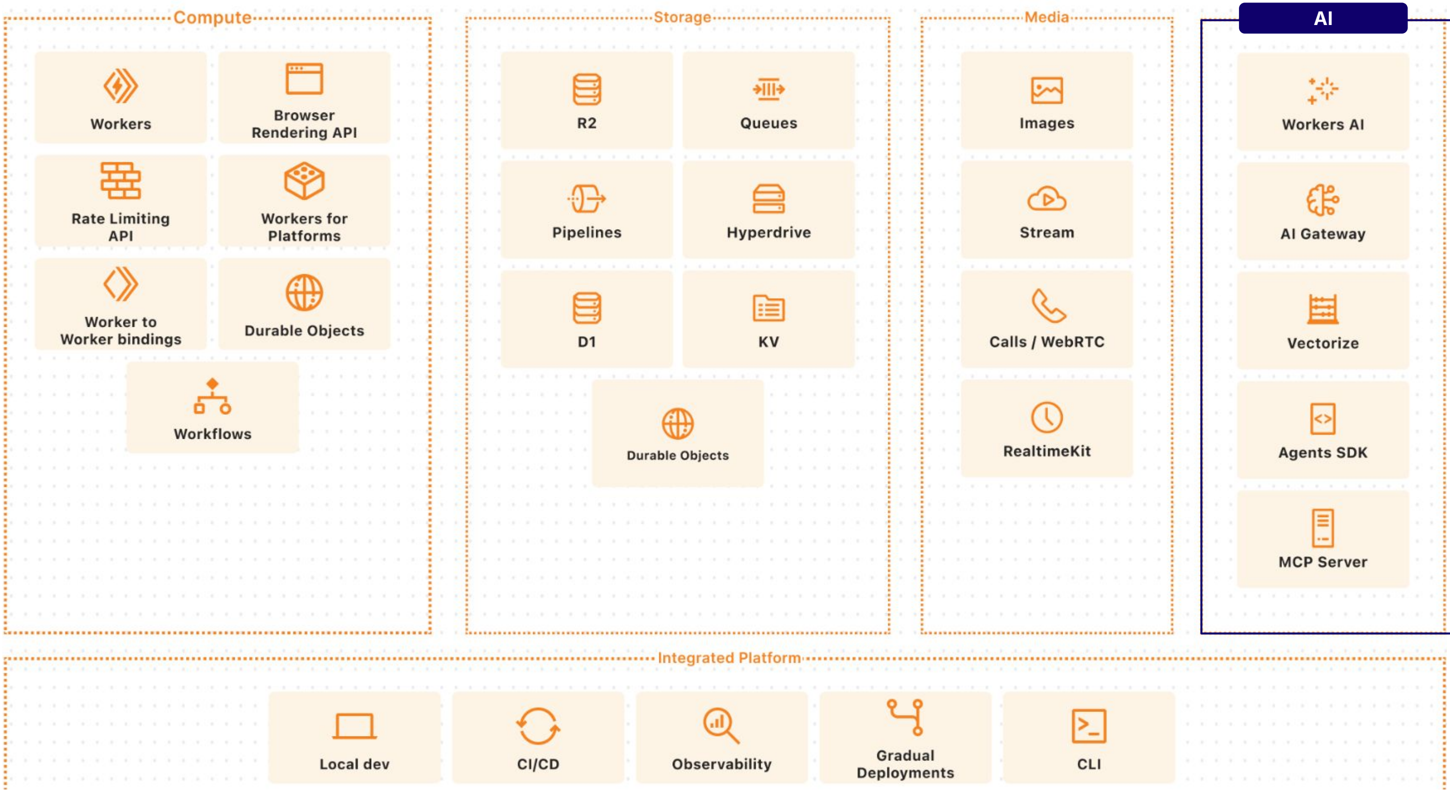
200+ cities

with GPUs

Growing constellation of cities for AI inference powered by GPUs



Cloudflare is the best developer platform to build Secure AI experiences



Workers AI
Global AI inference as a service

AI Gateway
Control and observe your AI applications

Vectorize
Globally distributed vector database

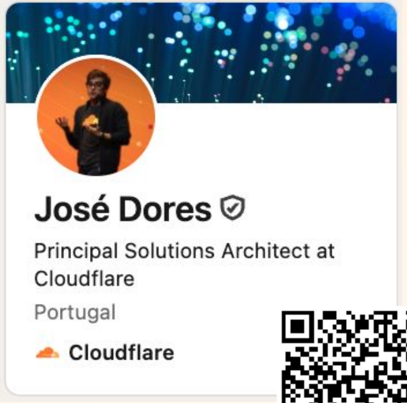
Agents SDK
Build and deploy AI-powered agents

MCP Server
Deploy remote MCP servers on Cloudflare

… and more

Thank you!

→ Let's connect!



José Dores

Principal Solutions Architect at Cloudflare

Portugal

Cloudflare

→ Want to dive deeper?

Accelerate your AI journey

Cloudflare's own MCP servers

→ Come visit us at the **Cloudflare booth**