CLOUDFLARE

EBOOK

Five ways DNS filtering helps your AI security strategy

Table of contents

Introduction

DNS filtering offers quick time-to-value for your AI security

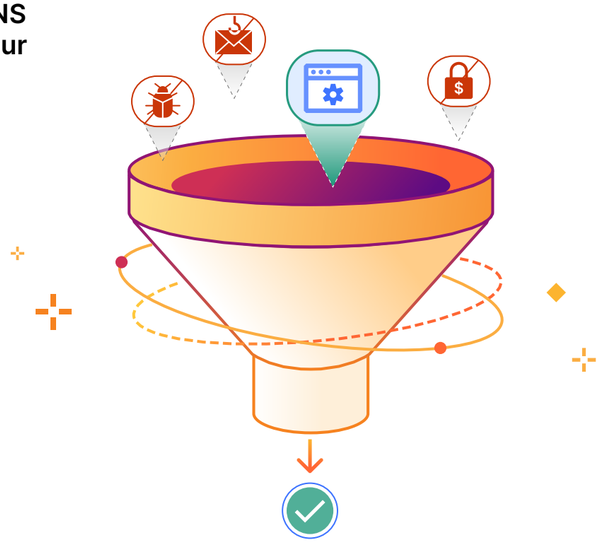
As organizations race to integrate artificial intelligence (AI) into their workflows, excitement to unlock productivity often hides growing security gaps. Uncontrolled use of generative AI tools like ChatGPT or Claude creates a lawless digital frontier where sensitive data is at risk and compliance falls by the wayside. At the same time, threat actors are weaponizing AI to supercharge their attacks and exploit this expanding attack surface.

Luckily, one of the most well-established security technologies, **DNS filtering**, can help businesses quickly embrace a more proactive, lightweight way to mitigate these risks.

DNS filtering — restricting web content based on domains and IPs — is traditionally seen as a simple and effective layer of protection to block Internet malware and enforce acceptable use policies. But it is also increasingly a popular early step for IT and security teams to begin modernizing their overall AI security approach.

This ebook highlights **five common ways DNS filtering with Cloudflare helps you adapt your security approach for the AI era:**

1. Discovering shadow AI
2. Controlling access to AI
3. Stopping AI-enhanced cyber threats
4. Preventing data exposure/exfiltration
5. Protecting AI development



From this initial foundation, organizations often deepen their visibility and controls across more environments, extending capabilities like HTTP inspections via a secure web gateway (SWG) or a broader secure access service edge (SASE) platform. This ebook also calls out how organizations can deploy SWG and SASE capabilities to further enhance their AI security approach:

Deployment phase with Cloudflare	Example capability
Step 1: Deploy DNS filtering	Analyze shadow AI use and enforce access controls based domains and IPs
Step 2: Deepen SWG inspections	Block user prompts in AI tools based on sensitive data detections and topical guardrails
Step 3: Extend SASE platform	Enforce AI usage controls across human-to-AI and machine-to-machine (agentic) communication.

1 Discover shadow AI/IT

Filter DNS queries for basic visibility

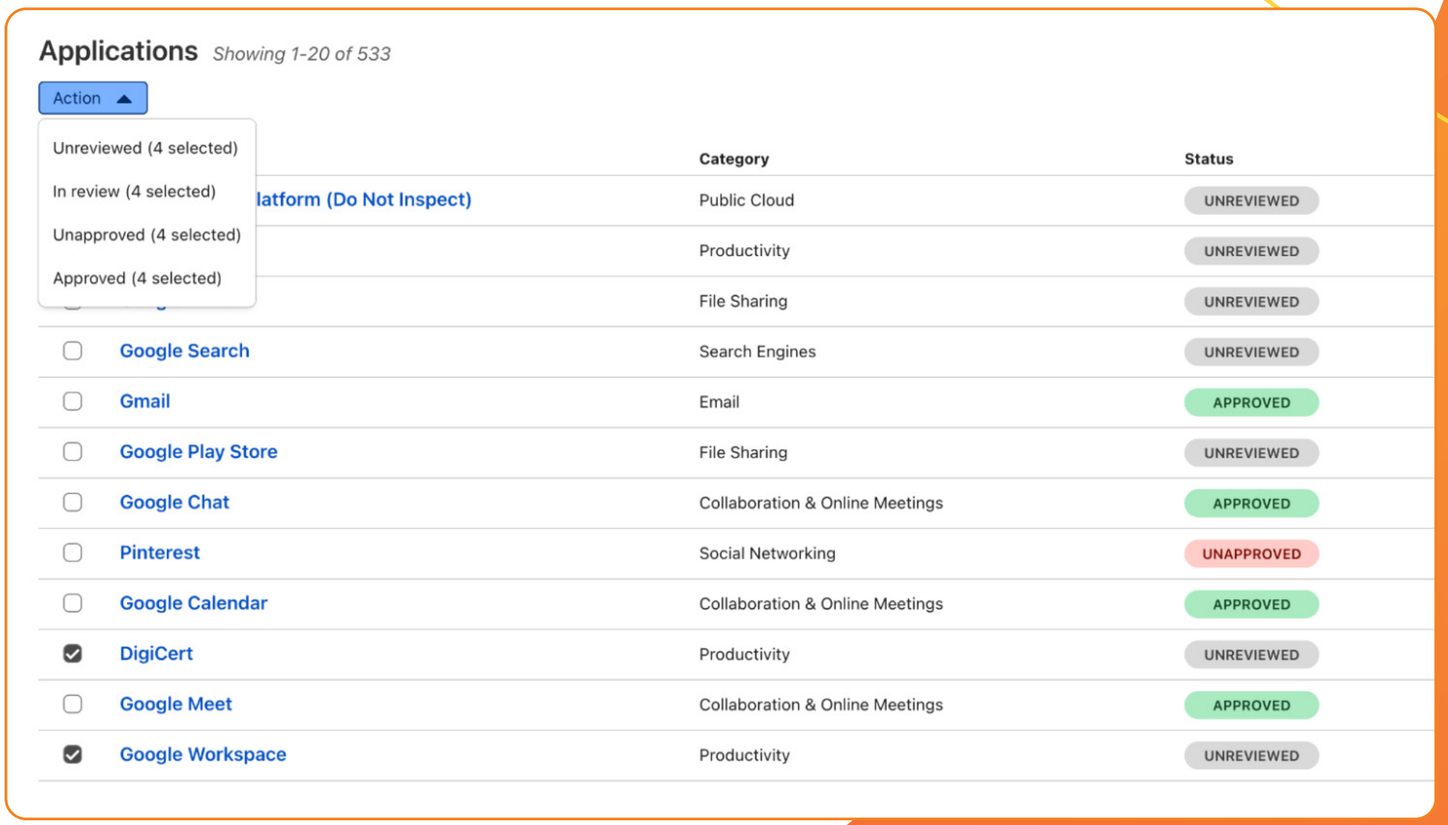
Organizations have dealt with unsanctioned / unapproved use of SaaS tools for years, but the explosion of AI tools and rush to use them is sparking today's shadow AI emergency:

20% of organizations suffered a breach due to incidents with shadow AI in 2025.[1]

85% of IT leaders say employees are adopting AI tools before IT can assess them.[2]

Filtering DNS queries helps you regain basic shadow AI visibility by tracking every DNS query made by your users. This allows you to:

- **Identify apps** based on domain resolution (ex. chatgpt.com or claude.ai)
- Categorize and review the **approval status of apps** based on domain (ex. approved, unapproved, unreviewed, or in review). See example to the right.
- Evaluate an app's trustworthiness based on **application confidence scores**. This score evaluates not only general risks posed by SaaS tools like compliance certifications and data management practices, but AI-specific risks including whether user data is used for model training or whether the model has a published system card detailing bias testing.



Review and mark application statuses in the dashboard

1 Discover shadow AI/IT

Beyond DNS Filtering

Refine access controls with HTTP policies

While DNS filtering provides a baseline for **who is going to what application,** turning on HTTP inspection allows for more granular views on **what they are doing within that app.** This visibility even includes logs of prompts and responses between users and generative AI tools.

Dashboards like the one below offer aggregated analytics for trends over time.

To analyze further, click on any AI app to see specific users or groups accessing it, their usage frequency, location, and more details.

Shadow IT analytics dashboard

One common line of inquiry is understanding data transfer patterns into and out of AI apps. **Shown below is a sample analysis of data downloaded/uploaded by hostname,** which can further filtered by user, content category, and other criteria.

Hostname by data downloaded

oneclient.sfx.ms	40.35 MB
www.bing.com	3.17 MB
chatgpt.com	3.14 MB
www.gstatic.com	2.17 MB
gemini.google.com	185.21 KB

Hostname by data uploaded

gemini.google.com	2.39 MB
play.google.com	399.97 KB
clients4.google.com	110.00 KB
go.microsoft.com	89.68 KB
www.bing.com	52.67 KB

2 Control access to AI

Set baseline access rules based on domain categories

DNS filtering is popular as a simple, lightweight way to restrict users from reaching malicious or unwanted Internet content. To protect their employees, organizations will typically block all domains and IPs automatically categorized as **security risks** like malware, phishing, command-and-control (C2) servers, botnet, and DNS tunneling destinations. They will also block **content categories** like adult, gambling, or video streaming as well as **specific classified apps.** This content filtering is often used to enforce acceptable use policies for employees or guests in shared spaces like a retail location, hotel, hospital, or school.

Use domain categories and app selectors to control which AI tools users can access. For example, combine two policies **to block all AI apps except one approved app, ChatGPT:**

Step 1 **Set ALLOW rule for ChatGPT**

See sample selector

Selector (Required)
Application

Operator (Required)
in

Value
ChatGPT

Step 2 **Set BLOCK rule for all other AI**

See sample selector

Selector (Required)
Content Categories

Operator (Required)
in

Value
Artificial Intelligence

DNS override actions even enable policies to redirect traffic bound for risky domains to specific internal resources or sinkhole servers based on IPs. For example, with Cloudflare:

Selector	Operator	Value	Action	Override
Hostname	is	www.riskyAI.com	Override	1.2.3.4 (internal AI policy page)

2 Control access to AI *cont.*

Beyond DNS Filtering

Refine access controls with HTTP policies

Turning on full proxy SWG inspection, organizations can enable more precise and flexible access controls with HTTP policies. Some popular approaches include:

- **Enforce policies for shadow AI based on app approval status:** Customize rules for apps approved/unapproved/unreviewed/in review. Blocking all unapproved AI apps is one direct option, but you can also apply more varied actions like the ones below.
- **Redirecting traffic to specific URLs:** For example, send user requests from unapproved AI tools to an approved one or an educational landing page.
- **Isolate session in a remote browser:** Route traffic for unreviewed, in review, or other specific apps in an isolated browser, where all web code runs on Cloudflare's network instead of on a local device. Isolation helps you protect data by controlling user actions, including restricting copy-paste, file uploads/downloads, keyboard inputs, and more.
- **Display custom notifications via device client:** Show a custom message via the Cloudflare device client when a user's traffic is blocked. This is often used to explain the rationale behind the block decision.

While these are common access policies, HTTP policies are needed for more granular data protection, including data loss prevention (DLP) detections, explored in the next section.

3 Stop AI-enhanced cyber threats and 4 Prevent data exposure/exfiltration

DNS filtering is still effective against emerging AI-driven threats and data theft

Threat actors are increasingly leveraging AI to execute, automate, and scale their attacks, often with the classic goal of exfiltrating sensitive data. These campaigns can be faster, more effective, and harder to detect:

76% of organizations admit they struggle to match the speed and sophistication of AI-powered attacks.[3]

Researchers have reported campaigns where AI performs **80-90%** of an attack, with only minimal human intervention required.[4]

While headlines tend to focus on novel AI techniques like deepfakes and polymorphic malware, attackers still rely on traditional methods and infrastructure. DNS filtering offers an effective first line of defense against both ends of that spectrum.

The table at right reflects common threats that DNS filtering services automatically block and how they power AI-enabled attacks to steal data. In particular, an authoritative and recursive DNS resolver like Cloudflare with real-time visibility across global Internet infrastructure (5.7+ trillion DNS queries per day) has unique telemetry to power threat hunting model to identify threats, often using AI and machine-learning (ML) to do so. In this way, security can proactively use AI to defend against AI.

Threat	Role in AI-enabled campaigns	How DNS filtering helps
Phishing domains	AI can generate hyper-personalized lures to drive targets to phishing domains, which often rely on "lookalike" domains (e.g., mybank-security.com instead of mybank.com). There, attackers can harvest credentials, steal sessions cookies, and worse.	Even if an employee clicks a phishing link, the request fails before the phishing page can load.
C2 callbacks	Even sophisticated AI-powered attacks aim to infect a device with malware. That malware still typically needs to "phone home" to a C2 server to receive further instructions.	Even if a device is already infected, DNS filtering can recognize and block queries sent to C2 servers to prevent it from executing its payload.
Newly seen and algorithmically generated domains	Attackers can use AI to generate unique, short-lived domains as infrastructure to bypass static blacklists and execute various stages of a campaign (e.g., C2 callbacks).	DNS filters classify and block queries to these domains. Vendors like Cloudflare with a high volume and frequency of DNS traffic excel in spotting these risks.
DNS tunneling	Attackers disguise data theft by encoding sensitive data into legitimate-looking DNS queries. AI can make it easier to mimic legitimate traffic and avoid detection in this encoding process, for example by transmitting queries at intervals that imitate human Internet browsing more closely.	DNS filters use AI and ML-backed models to analyze the mathematical, behavioral, and structural properties of DNS queries to detect and block tunneling attempts.

4 Prevent data exposure/exfiltration *cont.*

Beyond DNS Filtering

Turn on HTTP controls to safeguard data flows when users interact with AI tools

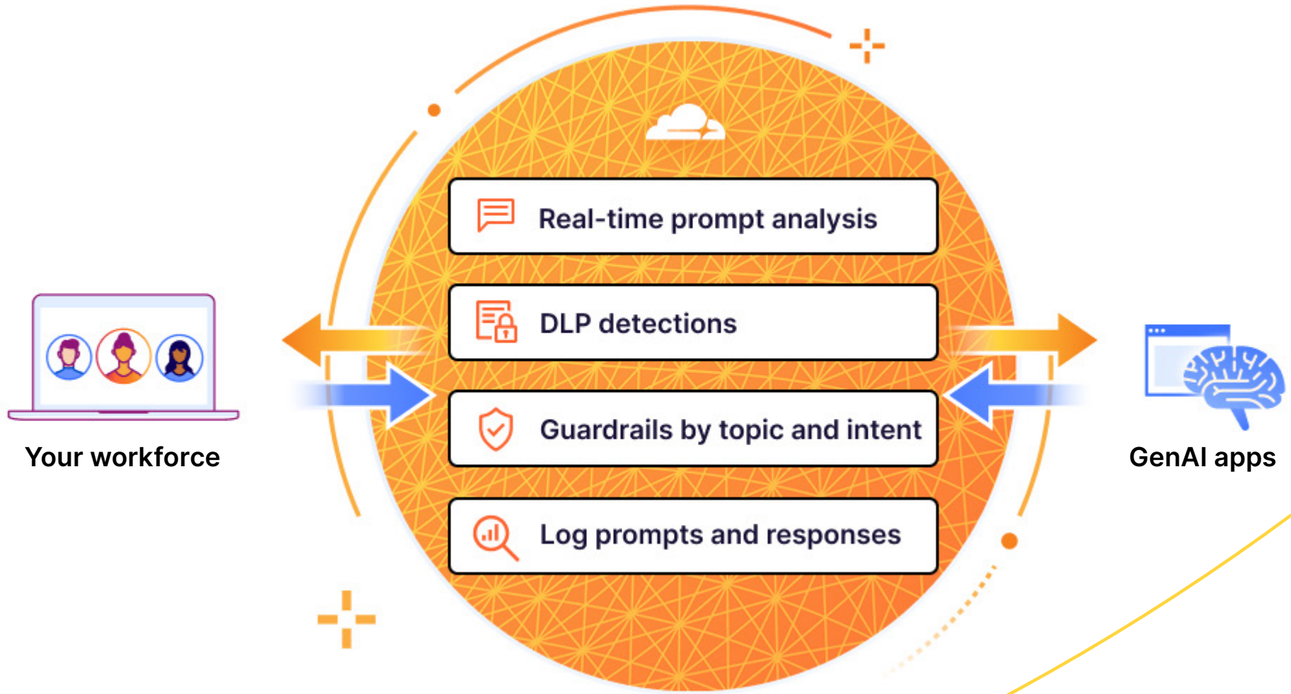
DNS filtering is efficient for its straightforward “allow or block” policies. But to encourage AI adoption, organizations want to move beyond this binary approach and, instead, focus on protecting data when users interact with AI tools.

With HTTP inspection turned on, a SWG like Cloudflare can detect, block, and log any user attempt to include sensitive data in an AI prompt. Here, Cloudflare uses classic **data loss prevention (DLP) detections** for personally identifiable information (PII), source code, customer data, financial information, credentials, and more.

The SWG can analyze not only the ***content,*** but also the ***context*** of AI prompts to stop data exposure. Here for example, Cloudflare looks for inappropriate and malicious purpose in a user prompt and creates **guardrails based on topic and intent** to prevent any risky outputs. So if a prompt tries to request PII or malicious code, or tries to get around an AI model’s policies, Cloudflare will block and log that too.

The reality is that most people will overshare data with AI. In fact, a recent survey found that **93% of employees** admit to putting info into AI tools without approval.[5] DLP detections and guardrails help security teams strike a balance between encouraging productivity and mitigating risks.

Real-time prompt protections and guardrails with Cloudflare’s SWG



5 Protect AI development

Protect developers building AI-enabled apps

More and more organizations are not only adopting AI tools, but also building their own AI-enabled apps in-house. Deploying DNS filtering protects developer teams responsible for building these AI experiences in their daily workflows. The same tried-and-true approach can mitigate new risks, including the ones below:

- **Blocking "model phishing" and data poisoning attempts:** AI apps rely heavily on external libraries, pre-trained models, and data from hubs like Hugging Face, and API integrations. Attackers can typosquat domains that host counterfeit AI services (for example, the "near-miss" *huggngface.co* instead of the correct *huggingface.co*). Developers can inadvertently reach these fake destinations by mistyping or clicking a link. There, they can be tricked into entering API credentials, downloading malicious code, or using poisoned models and datasets. **A DNS filter would intercept and block queries to this risky and often newly registered domains, preventing these phishing and supply chain attacks.**
- **Stop model weight exfiltration:** The weights of an AI model are its crown jewels, representing high-value intellectual property. A common exfiltration tactic is to use a compromised developer machine to upload these files to obscure file-sharing domains or private repositories. **The right DNS filtering policies (e.g., restrict DNS resolution only to sanctioned resources) would block the machine's request before any data transfer even begins.**
- **Block indirect prompt injections:** A developer may task an AI agent to parse a webpage or content that contains hidden instructions with malicious intent. Those instructions might tell the AI to fetch more data from a specific domain or run a C2 callback to a compromised server. **DNS filters can prevent this indirect prompt injection by preventing the agent's data pulldown or phone home attempts.**

Spotlight on Cloudflare's developer platform

Build world-class, secure AI experiences

Cloudflare's developer platform provides the infrastructure to scale your AI applications at every step — build AI apps and agents, store training data, run AI inference — with security by design.

- **Control and observe AI-powered apps,** while reducing inference costs and dynamically routing traffic
- **Build model context protocol (MCP) servers** with authentication and authorization built in
- **Prevent service disruptions** with model fallbacks and rate limiting

Looking ahead: Securing AI adoption with Cloudflare One

Start with DNS filtering

As a standalone solution, DNS filtering offers a simple, effective way to navigate key challenges and opportunities of AI. Deployments with or without a device client and intuitive policy management help security and IT teams start realizing value quickly across hybrid workforces.

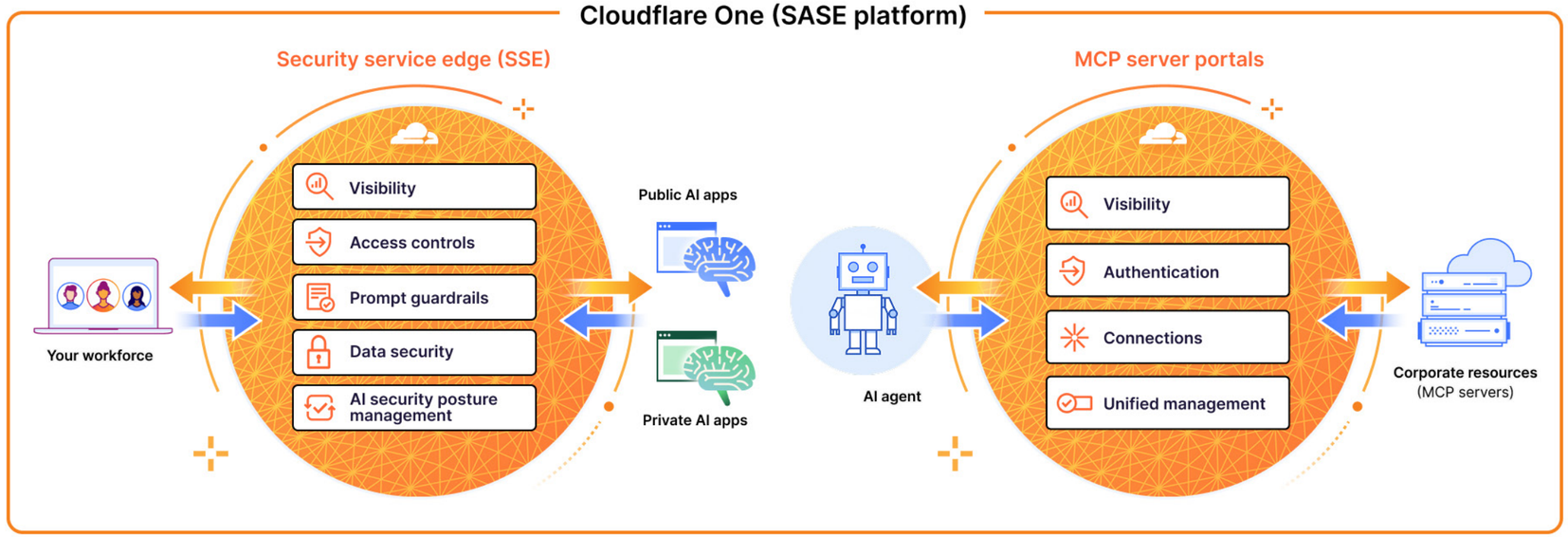
For many organizations, modernizing DNS filtering is a common early step toward a full SWG deployment or consolidated SASE architecture. Platforms like Cloudflare that streamline this progression position your organizations to adapt with agility and safely accelerate AI adoption.

Beyond DNS Filtering

Extend SWG and SASE platform to secure workforce use of generative and agentic AI

Cloudflare One, a SASE platform, sits in between your workforce and AI tools, making it a logical control point to protect use of AI tools. Whether employees are chatting with ChatGPT or AI agents are gathering information across corporate resources, Cloudflare offers consistent visibility and security across both human-to-AI and machine-to-machine interactions — all from one unified dashboard and control plane:

- **Discover shadow AI** and manage policies for all sanctioned and unsanctioned AI tools
- **Strengthen AI governance** with identity-based access controls for GenAI use and agentic AI communication
- **Stop data loss** by blocking sensitive information in user prompts, enforcing topical guardrails, and scanning for misconfigurations in AI



References

1. 2025 IBM, Cost of a Data Breach report: https://newsroom.ibm.com/2025-07-30-ibm-report-13-of-organizations-reported-breaches-of-ai-models-or-applications,-97-of-which-reported-lacking-proper-ai-access-controls
2. 2025 ManageEngine research: https://www.manageengine.com/survey/shadow-ai-surge-enterprises/
3. CrowdStrike Ransomware Report: AI Attacks Outpacing Defenses, 2025: https://www.crowdstrike.com/en-us/press-releases/ransomware-report-ai-attacks-outpacing-defenses/
4. “Disrupting the first reported AI-orchestrated cyber espionage campaign,” Anthropic, Nov, 13, 2025: https://www.anthropic.com/news/disrupting-AI-espionage
5. 2025 ManageEngine research: https://www.manageengine.com/survey/shadow-ai-surge-enterprises/

1 888 99 FLARE | enterprise@cloudflare.com | Cloudflare.com

REV:BDES-8725.2026JAN28