

Zero Trust in the Era of Artificial Intelligence

Sheril Nagoor
Principal Solutions Architect

*TAAFT for short

THERE'S AN AI FOR THAT*

15,385 AIs for  14,447 tasks and 4,803 jobs.

Spotlight: [Tweet Hunter \(Twitter management\)](#)

|

Find AIs using AI

/



#1 AI aggregator. Updated daily. Used by 20M+ humans.

43%

of employees have used AI-powered tools for work tasks — and over 2/3 did so without telling their boss

(Source: Fishbowl)



The current state of AI feels like an Ox. It is stronger than a human and can pull a plough but needs either constant human guidance or narrow fencing.

It is not quite at the level where we can ride it like a horse

–Joscha Bach



Picture Generated by <https://ai.cloudflare.com/>

The Challenges of AI Adoption

- Security concerns are a major barrier to AI adoption
- Organizations are hesitant to trust AI systems with sensitive data
- AI systems can be vulnerable to hacking, privacy breaches and other cyber attacks including injections or data exfiltration
- It is expensive to run a model

BREAKING

Samsung Bans ChatGPT Among Employees After Sensitive Code Leak

Siladitya Ray Forbes Staff

Covering breaking news and tech policy stories at Forbes.

[Follow](#)

Updated May 2, 2023, 07:31am EDT

f **TOPLINE** Samsung Electronics has banned the use of ChatGPT and other AI-powered chatbots by its employees, Bloomberg [reported](#), becoming the latest company to crack down on the workplace use of AI services amid concerns about sensitive internal information being leaked on such platforms.

X

in



Support us →

The
Guardian



Artificial intelligence (AI)

AI will make scam emails look genuine, UK cybersecurity agency warns

NCSC says generative AI tools will soon allow amateur cybercriminals to launch sophisticated phishing attacks

Dan Milmo *Global technology editor* and **Alex Hern**

Wed 24 Jan 2024 01:01 CET

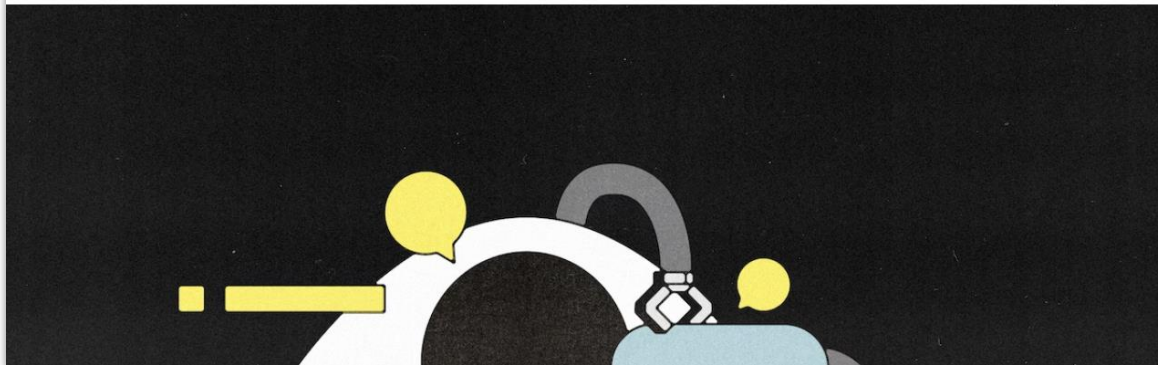
Chatbots are so gullible, they'll take directions from hackers

'Prompt injection' attacks haven't caused giant problems yet. But it's a matter of time, researchers say.



By [Tatum Hunter](#)

Updated November 2, 2023 at 3:13 p.m. EDT | Published November 2, 2023 at 6:00 a.m. EDT



OWASP Top 10 for LLM Applications

LLM01

Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02

Insecure Output Handling

This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.

LLM03

Training Data Poisoning

This occurs when LLM training data is tampered, introducing vulnerabilities or biases that compromise security, effectiveness, or ethical behavior. Sources include Common Crawl, WebText, OpenWebText, & books.

LLM04

Model Denial of Service

Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.

LLM05

Supply Chain Vulnerabilities

LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre-trained models, and plugins can add vulnerabilities.

LLM06

Sensitive Information Disclosure

LLMs may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. It's crucial to implement data sanitization and strict user policies to mitigate this.

LLM07

Insecure Plugin Design

LLM plugins can have insecure inputs and insufficient access control. This lack of application control makes them easier to exploit and can result in consequences like remote code execution.

LLM08

Excessive Agency

LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.

LLM09

Overreliance

Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

LLM10

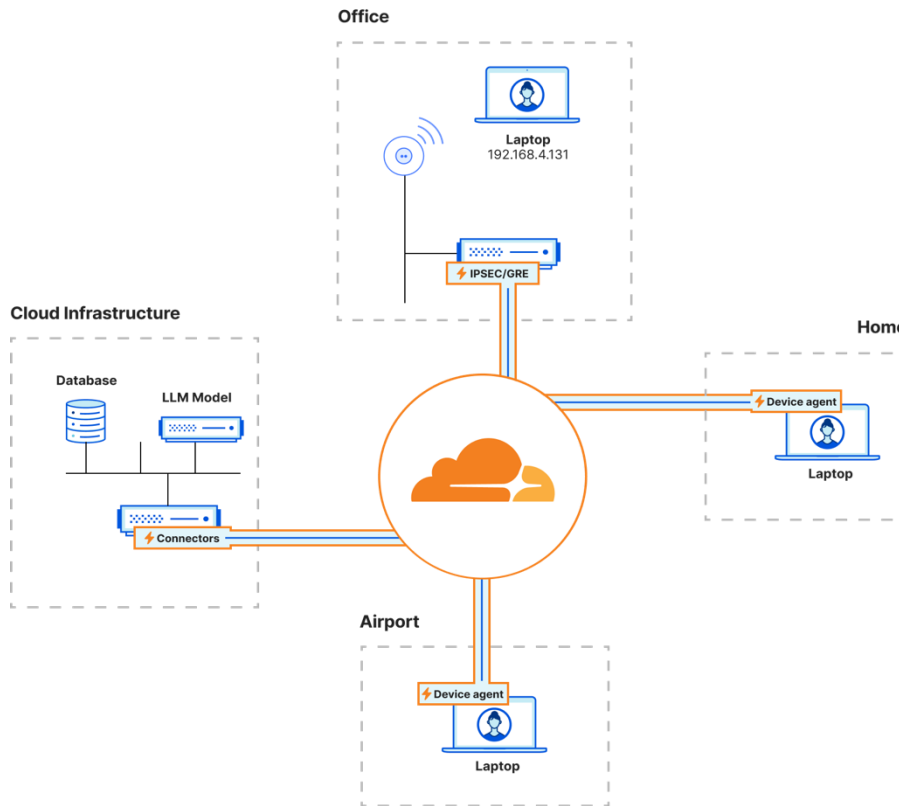
Model Theft

This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.

**Goal is to allow safe use the of AI tools, wherever they live,
without compromising performance.**

No user or device is trusted by default and all access must be continuously verified

Cloudflare One for AI















How to securely use AI

- Measure usage
- Control API access
- Restrict data uploads
- Check for misconfigurations*



Measure Usage

- Employees Testing
- Data Control
- Approved Tool

<input type="checkbox"/>	Application	Application Type	Status	Secured
<input type="checkbox"/>	 Stripe	Finance & Accounting	 Unreviewed	No
<input type="checkbox"/>	 Bing	Search Engines	 Unreviewed	No
<input type="checkbox"/>	 DuckDuckGo	Search Engines	 Unreviewed	No
<input type="checkbox"/>	 ChatGPT	Artificial Intelligence	 In Review	No
<input type="checkbox"/>	 Microsoft Copilot	Artificial Intelligence	 Unreviewed	No
<input type="checkbox"/>	 Google Gemini	Artificial Intelligence	 Unreviewed	No

Control API access

- Control and scoped Inputs
- Securely share training data
- Tokens for each model and systems making API requests
- Log every requests
- Ability to revoke as needed

[← Back to Service Tokens](#)

AI Model - HelperBot

Service token name (required)
AI Model - HelperBot

Service Token Duration (required)
1 year

Service token details

Header and client ID
CF-Access-Client-Id: 3b836bac3db4ea6968412a3226713543, access

Header and client secret
CF-Access-Client-Secret: [REDACTED]

*Client secret is only displayed during the creation of the service token. Save it in a secure place.

Created at
September 4, 2024 • 2:14 PM
2 seconds ago

Configure rules

The rules you create here define who can or cannot reach your application.

Include

Selector	Value
Service Token	AI Model - HelperBot

Require

Selector	Value
Country	Estonia

Selector

Value
103.220.220.10 103.220.220.20

IPv4/v6 addresses and CIDRs

Restrict Data Uploads

- Stop oversharing of sensitive data
- Avoid security incidents
- Use predefined data set
- Create custom data set
- Control who is allowed to experiment AI tools

DLP / DLP profiles

Data Loss Prevention profiles

Configure Data Loss Prevention (DLP) profiles to scan uploaded or downloaded files for sensitive data. To allow or disallow file transfer, apply [Gateway HTTP policies](#).

[DLP profile documentation](#)

Your DLP profiles

To use Microsoft Information Protection (MIP) sensitivity labels, add your MIP account through CASB integration. [Learn more](#)

[+ Create profile](#)

Search by profile name

Profile name	Profile type
Credentials and Secrets	PRE-DEFINED
Financial Information	PRE-DEFINED
Health Information	PRE-DEFINED
HelperBot	CUSTOM
Social Security, Insurance, Tax, and Identifier Numbers	PRE-DEFINED
Source Code	PRE-DEFINED

Traffic

Selector (Required) Operator (Required) Value

Application in Artificial Intelligence Google Bard Chat & Ask AI ChatGPT Microsoft Copilot Databot
Elsa Speak Google Gemini Lensa Otter Seeing AI synthesia.io

And

Selector (Required) Operator (Required) Value

DLP Profile in Credentials and Secrets HelperBot Source Code

+ And

+ Or

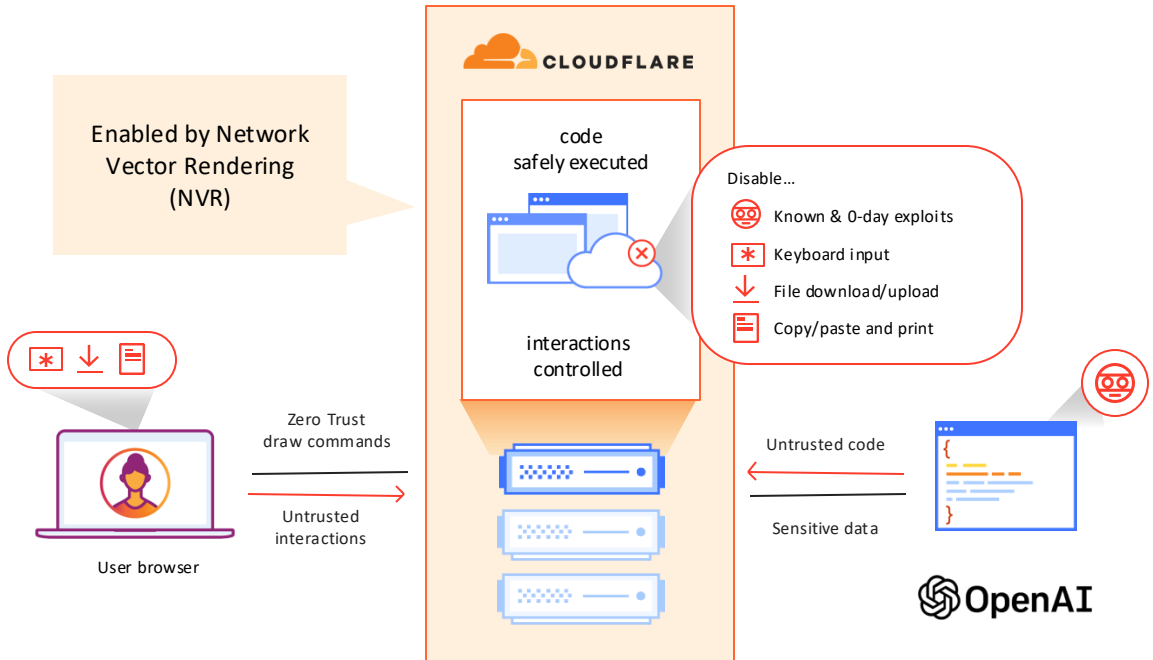
Identity

Selector (Required) Operator (Required) Value

User Group Names in AI Research Team

+ And

Allowing, but limiting, chatbot usage



Check for Misconfigurations*

- Integrate with popular AI services
- Check for misconfiguration before it turns to a security incident



67%

of organizations have adopted some form of AI in their **software development** processes.

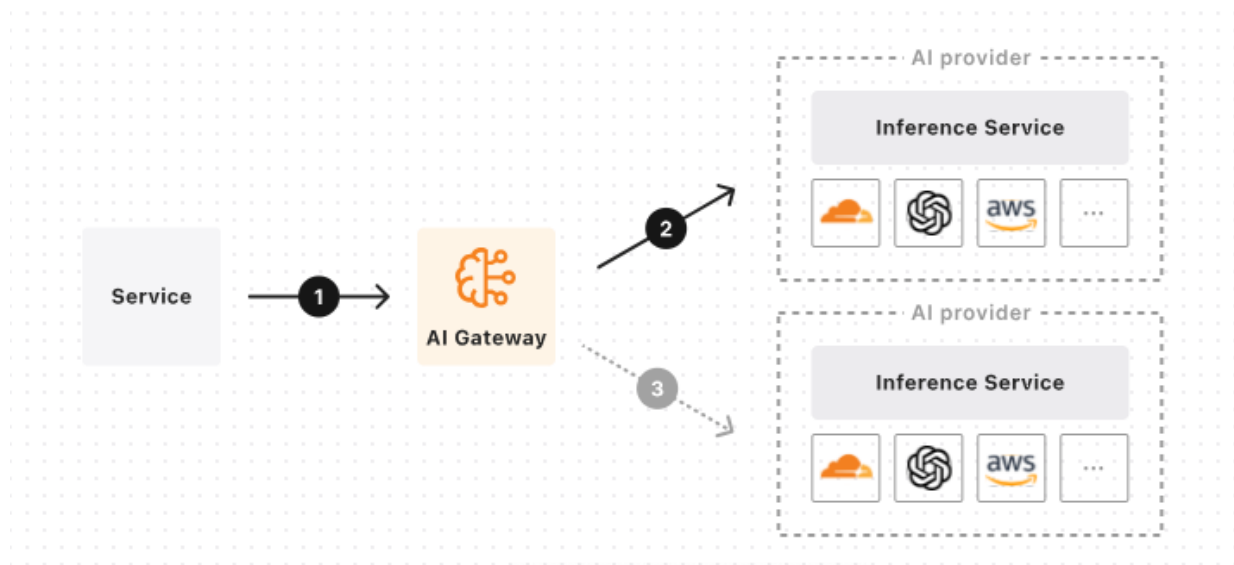
(Source: McKinsey)



Secure your applications that are using LLM

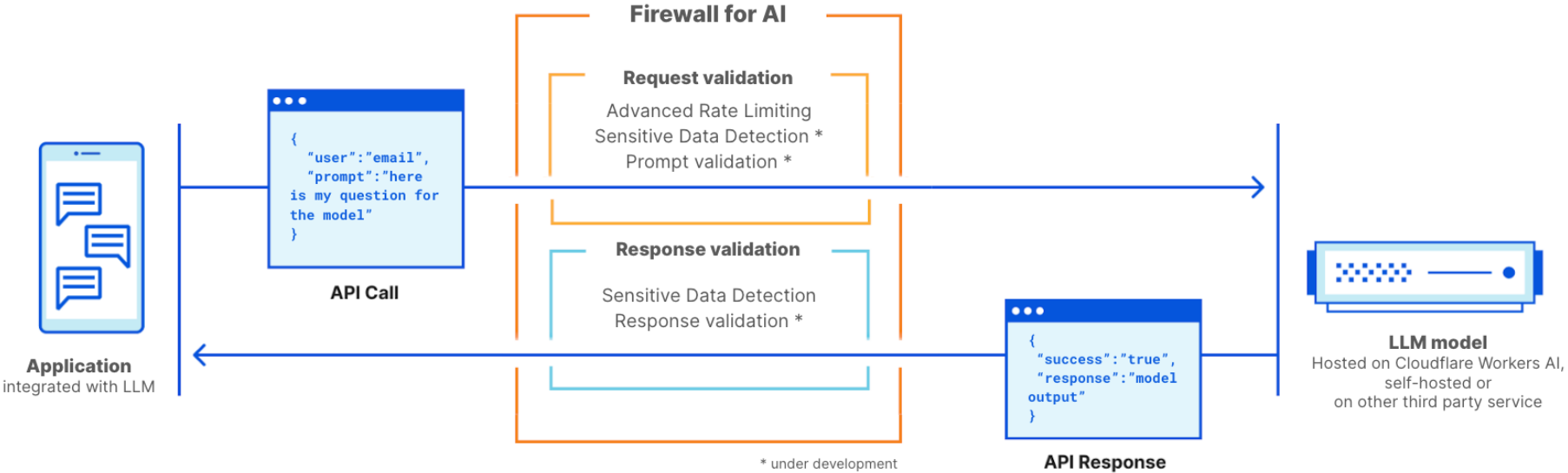
- Gain Visibility
- Identify abuses before they reach the models
- Detect sensitive data exfiltration from the model

AI Gateway - Observe and control AI usage



- Gain visibility and control over your AI apps
- Insights into caching, rate limiting, as well as request - retries, model fallback, and more
- Only takes one line of code to get started.

Firewall for AI



Start today!



Cloudflare's connectivity cloud

Composable, programmable
architecture

Integration with
AI networks

Platform intelligence
and innovations

Simple, unified
interface

Connect

- SASE
- Apps
- Network Interconnect
- Smart Routing
- + More

Protect

- SSE
- Email Security
- WAF/API Security
- DDoS (L3 & L7)
- Network Security
- + More

Build

- Serverless AI
- AI Gateway
- Object Storage
- Video Streaming
- + More

One programmable global cloud network

Let's Connect



Sheril Nagoor

Driving Cybersecurity Excellence: Empowering
Organizations to Embrace Zero Trust and SASE S...

