CLOUDFLARE

EBOOK

Cinque modi in cui il filtro DNS contribuisce a migliorare la tua strategia di sicurezza IA

Indice

Introduzione

Il filtro DNS offre un time-to-value rapido per la tua sicurezza IA

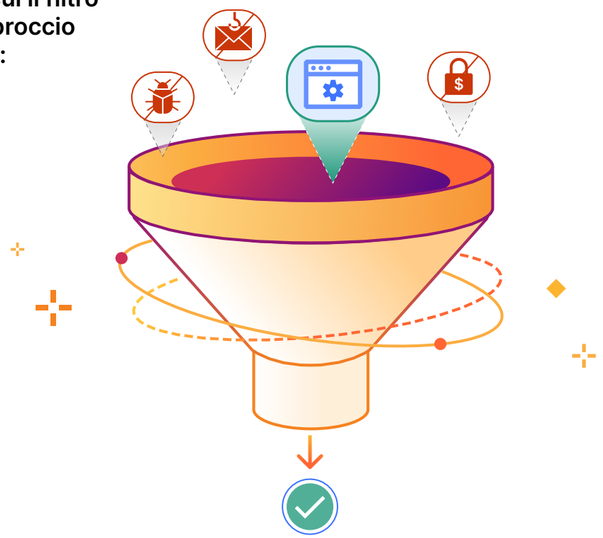
Mentre le organizzazioni corrono per integrare l'intelligenza artificiale (IA) nei loro flussi di lavoro, l'entusiasmo per sbloccare la produttività spesso nasconde crescenti lacune nella sicurezza. L'uso incontrollato di strumenti di IA generativa come ChatGPT o Claude crea una frontiera digitale senza legge in cui i dati sensibili sono a rischio e la compliance cade nel dimenticatoio. Allo stesso tempo, gli autori delle minacce usano l'IA per potenziare i loro attacchi e sfruttare questa superficie d'attacco in espansione.

Fortunatamente, una delle tecnologie di sicurezza più consolidate, il **filtro DNS**, può aiutare le aziende ad adottare rapidamente un modo più proattivo e leggero per mitigare questi rischi.

Il filtro DNS, che limita i contenuti web in base a domini e IP, è tradizionalmente visto come un livello di protezione semplice ed efficace per bloccare il malware su Internet e applicare criteri di utilizzo accettabili. Ma è anche un primo passo sempre più popolare per i team IT e di sicurezza iniziare a modernizzare il loro approccio generale alla sicurezza dell'IA.

Questo ebook evidenzia **cinque modi comuni in cui il filtro DNS con Cloudflare ti aiuta ad adattare il tuo approccio alla sicurezza per l'era dell'intelligenza artificiale:**

1. Scopri la shadow AI
2. Controllo dell'accesso all'IA
3. Bloccare le minacce informatiche potenziate dall'intelligenza artificiale
4. Prevenzione dell'esposizione/esfiltrazione dei dati
5. Protezione dello sviluppo dell'IA



Da questa base iniziale, le organizzazioni spesso approfondiscono la propria visibilità e i propri controlli su più ambienti, estendendo funzionalità come le ispezioni HTTP tramite un gateway web sicuro (SWG) o una piattaforma SASE (Secure Access Service Edge) più ampia. Questo ebook spiega anche come le organizzazioni possono distribuire le funzionalità SWG e SASE per migliorare ulteriormente il loro approccio alla sicurezza dell'IA:

Fase di distribuzione con Cloudflare	Capacità di esempio
Passaggio 1: distribuisci il filtro DNS	Analizza l'uso della shadow AI e applica i controlli di accesso basati su domini e IP
Passaggio 2: approfondisci le ispezioni SWG	Blocca i prompt degli utenti negli strumenti IA in base ai rilevamenti di dati sensibili e ai vincoli tematici
Fase 3: estendi la piattaforma SASE	Applica i controlli sull'utilizzo dell'intelligenza artificiale nella comunicazione da uomo a intelligenza artificiale e da macchina a macchina (agentica).

1 Scopri shadow AI e IT

Filtra le query DNS per una visibilità di base

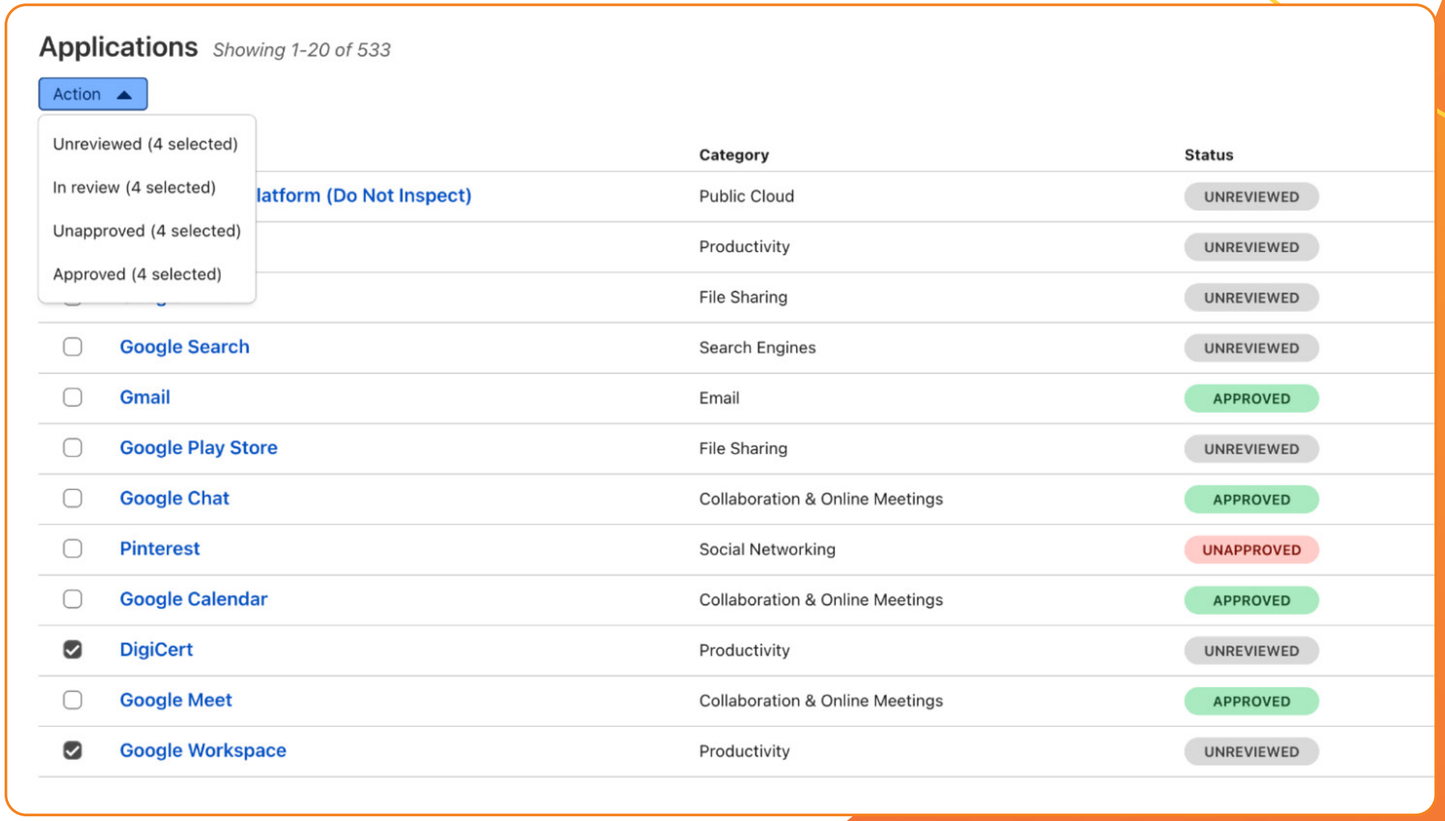
Le organizzazioni hanno affrontato per anni l'uso non autorizzato/non approvato degli strumenti SaaS, ma l'esplosione degli strumenti IA e la fretta di usarli sta innescando l'odierna emergenza della shadow AI:

Il 20% delle organizzazioni ha subito una violazione a causa di incidenti di sicurezza con la shadow AI nel 2025.[1]

L'85% dei responsabili IT afferma che i dipendenti stanno adottando strumenti IA prima che l'IT possa valutarli.[2]

Il filtraggio delle query DNS ti aiuta a riguadagnare la visibilità di base della shadow AI monitorando ogni query DNS eseguita dai tuoi utenti. Ciò consente di:

- **Identificare le app** in base alla risoluzione del dominio (ad es. chatgpt.com o claude.ai)
- Classificare e rivedere lo **stato di approvazione delle app** in base al dominio (ad es. approvato, non approvato, non rivisto o in revisione). Vedi l'esempio a destra.
- Valutare l'affidabilità di un'app in base a **punteggi di attendibilità dell'applicazione**. Questo punteggio valuta non solo i rischi generali posti dagli strumenti SaaS come le certificazioni di compliance e le pratiche di gestione dei dati, ma anche i rischi specifici dell'IA, incluso se i dati utente vengono utilizzati per l'addestramento del modello o se il modello ha una scheda di sistema pubblicata in cui sono indicati nel dettaglio i test di distorsione.



Rivedi e contrassegnar gli stati delle applicazioni nel dashboard

1 Scopri shadow AI e IT

Oltre il filtro DNS

Perfeziona i controlli di accesso con i criteri HTTP

Mentre il filtro DNS fornisce una linea di base per sapere **chi vuole utilizzare una determinata applicazione,** l'attivazione dell'ispezione HTTP consente visualizzazioni più granulari su **le azioni compiute in tale app.** Questa visibilità include anche i registri dei prompt e delle risposte tra gli utenti e gli strumenti di IA generativa.

I dashboard come quello di seguito offrono analisi aggregate per le tendenze nel tempo.

Per analizzare ulteriormente, fai clic su qualsiasi app IA per vedere utenti o gruppi specifici che accedono ad essa, la loro frequenza di utilizzo, posizione e maggiori dettagli.

Dashboard di analisi dello shadow IT

Una linea di indagine comune è la comprensione dei modelli di trasferimento dei dati dentro e fuori le app IA. **Di seguito è mostrata un'analisi di esempio dei dati scaricati/caricati per nome host,** che possono essere ulteriormente filtrati per utente, categoria di contenuto e altri criteri.

Nome host in base ai dati scaricati

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

Nome host in base ai dati caricati

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 Controla l'accesso all'IA

Imposta le regole di accesso di base in base alle categorie di dominio

Il filtro DNS è popolare come un modo semplice e leggero per impedire agli utenti di raggiungere contenuti Internet dannosi o indesiderati. Per proteggere i propri dipendenti, le organizzazioni in genere bloccheranno tutti i domini e gli IP classificati automaticamente come **rischi per la sicurezza**, ad esempio malware, phishing, server di comando e controllo (C2), botnet e destinazioni di tunneling DNS. Bloccheranno anche **categorie di contenuti** come contenuti per adulti, giochi d'azzardo o streaming video, nonché **app classificate specifiche.** Questo filtro dei contenuti viene spesso utilizzato per applicare criteri di utilizzo accettabile per dipendenti o ospiti in spazi condivisi come un punto vendita, un hotel, un ospedale o una scuola.

Utilizza le categorie di dominio e i selettori di app per controllare a quali strumenti IA possono accedere gli utenti. Ad esempio, combina due criteri per **bloccare tutte le app di IA tranne un'app approvata, ChatGPT:**

Passo 1

Imposta la regola ALLOW per ChatGPT

Vedi il selettore di esempio

Selector (Required)
Application

Operator (Required)
in

Valore
ChatGPT

Passo 2

Imposta la regola BLOCK per tutte le altre IA

Vedi il selettore di esempio

Selector (Required)
Categorie di contenuti

Operator (Required)
in

Valore
Artificial Intelligence

Le **azioni di override del DNS** consentono inoltre ai criteri di reindirizzare il traffico destinato a domini rischiosi a risorse interne specifiche o server sinkhole basati su IP. Ad esempio, con Cloudflare:

Selettore	Operatore	Valore	Azione	Override
Nome host	è	www.riskyAI.com	Override	1.2.3.4 (pagina dei criteri IA interni)

2 Controlla l'accesso all'IA *cont.*

Oltre il filtro DNS

Perfeziona i controlli di accesso con i criteri HTTP

Attivando l'ispezione SWG completa del proxy, le organizzazioni possono abilitare controlli degli accessi più precisi e flessibili con i criteri HTTP. Alcuni approcci popolari includono:

- **Applicazione dei criteri per la shadow AI in base allo stato di approvazione delle app:** personalizza le regole per le app approvate/non approvate/non riviste/in revisione. Il blocco di tutte le app IA non approvate è un'opzione diretta, ma puoi anche applicare azioni più varie come quelle di seguito.
- **Reindirizzamento del traffico a URL specifici:** ad esempio, invia le richieste degli utenti da strumenti IA non approvati a uno approvato o a una pagina di destinazione didattica.
- **Isolamento della sessione in un browser remoto:** instrada il traffico per app non riviste, in revisione o altre app specifiche in un browser isolato, in cui tutto il codice web viene eseguito sulla rete di Cloudflare anziché su un dispositivo locale. L'isolamento ti aiuta a proteggere i dati controllando le azioni dell'utente, inclusa la limitazione di copia-incolla, caricamenti/download di file, input da tastiera e altro ancora.
- **Visualizzazione di notifiche personalizzate tramite il client del dispositivo:** mostra un messaggio personalizzato tramite il client del dispositivo Cloudflare quando il traffico di un utente è bloccato. Questo viene spesso utilizzato per spiegare la logica alla base della decisione di blocco.

Sebbene si tratti di criteri di accesso comuni, i criteri HTTP sono necessari per una protezione dei dati più granulare, inclusi i rilevamenti di prevenzione della perdita di dati (DLP), descritti nella sezione successiva.

3 Blocca le minacce informatiche potenziate dall'intelligenza artificiale e 4 previeni l'esposizione/esfiltrazione dei dati

Il filtro DNS è ancora efficace contro le minacce emergenti basate sull'intelligenza artificiale e il furto di dati

Gli autori delle minacce sfruttano sempre più l'intelligenza artificiale per eseguire, automatizzare e scalare i propri attacchi, spesso con il classico obiettivo di esfiltrare dati sensibili. Queste campagne possono essere più veloci, più efficaci e più difficili da rilevare:

Il **76%** delle organizzazioni ammette di faticare a eguagliare la velocità e la sofisticatezza degli attacchi basati sull'IA.[3]

I ricercatori hanno segnalato campagne in cui l'intelligenza artificiale esegue l'**80-90%** di un attacco, con un intervento umano minimo richiesto.[4]

Sebbene i titoli dei giornali tendano a concentrarsi su nuove tecniche di intelligenza artificiale come i deepfake e il malware polimorfico, gli aggressori si affidano ancora a metodi e infrastrutture tradizionali. Il filtro DNS offre un'efficace prima linea di difesa contro entrambe le estremità di questo spettro.

La tabella a destra riflette le minacce comuni che i servizi di filtro DNS bloccano automaticamente e il modo in cui alimentano gli attacchi basati sull'intelligenza artificiale per sottrarre dati. In particolare, un resolver DNS autoritativo e ricorsivo come Cloudflare con visibilità in tempo reale sull'infrastruttura Internet globale (oltre 5,7 trilioni di query DNS al giorno) dispone di una telemetria unica per potenziare il modello di caccia alle minacce per identificare le minacce, spesso utilizzando l'intelligenza artificiale e il machine learning (ML) per farlo. In questo modo, la sicurezza può utilizzare in modo proattivo l'IA per difendersi dall'IA.

Minaccia	Ruolo nelle campagne basate sull'intelligenza artificiale	Come aiuta il filtro DNS
Domini di phishing	L'intelligenza artificiale può generare esche iper-personalizzate per indirizzare gli obiettivi verso domini di phishing, che spesso si basano su domini "simili" (ad esempio, mybank-security.com invece di mybank.com). Lì, gli aggressori possono raccogliere credenziali, sottrarre cookie di sessione e compiere azioni ancora più nefaste.	Anche se un dipendente fa clic su un collegamento di phishing, la richiesta non riesce prima che la pagina di phishing possa essere caricata.
Callback C2	Anche gli attacchi sofisticati basati sull'intelligenza artificiale mirano a infettare un dispositivo con malware. Quel malware in genere ha ancora bisogno di "telefonare a casa" a un server C2 per ricevere ulteriori istruzioni.	Anche se un dispositivo è già infetto, il filtro DNS può riconoscere e bloccare le query inviate ai server C2 per impedirgli di eseguire il suo payload dannoso.
Domini appena visti e generati algoritmicamente	Gli autori di attacchi possono utilizzare l'intelligenza artificiale per generare domini univoci e di breve durata come infrastruttura per aggirare le blacklist statiche ed eseguire varie fasi di una campagna (ad esempio, callback C2).	I filtri DNS classificano e bloccano le query a questi domini. Fornitori come Cloudflare con un volume e una frequenza elevati di traffico DNS eccellono nell'individuare questi rischi.
Tunneling DNS	Gli autori di attacchi mascherano il furto di dati codificando i dati sensibili in query DNS dall'aspetto legittimo. L'intelligenza artificiale può semplificare l'imitazione del traffico legittimo ed evitare il rilevamento in questo processo di codifica, ad esempio trasmettendo query a intervalli che imitano più da vicino la navigazione umana su Internet.	I filtri DNS utilizzano modelli supportati da intelligenza artificiale e ML per analizzare le proprietà matematiche, comportamentali e strutturali delle query DNS per rilevare e bloccare i tentativi di tunneling.

4 Impedisci l'esposizione/esfiltrazione dei dati *cont.*

Oltre il filtro DNS

Attiva i controlli HTTP per salvaguardare i flussi di dati quando gli utenti interagiscono con gli strumenti IA

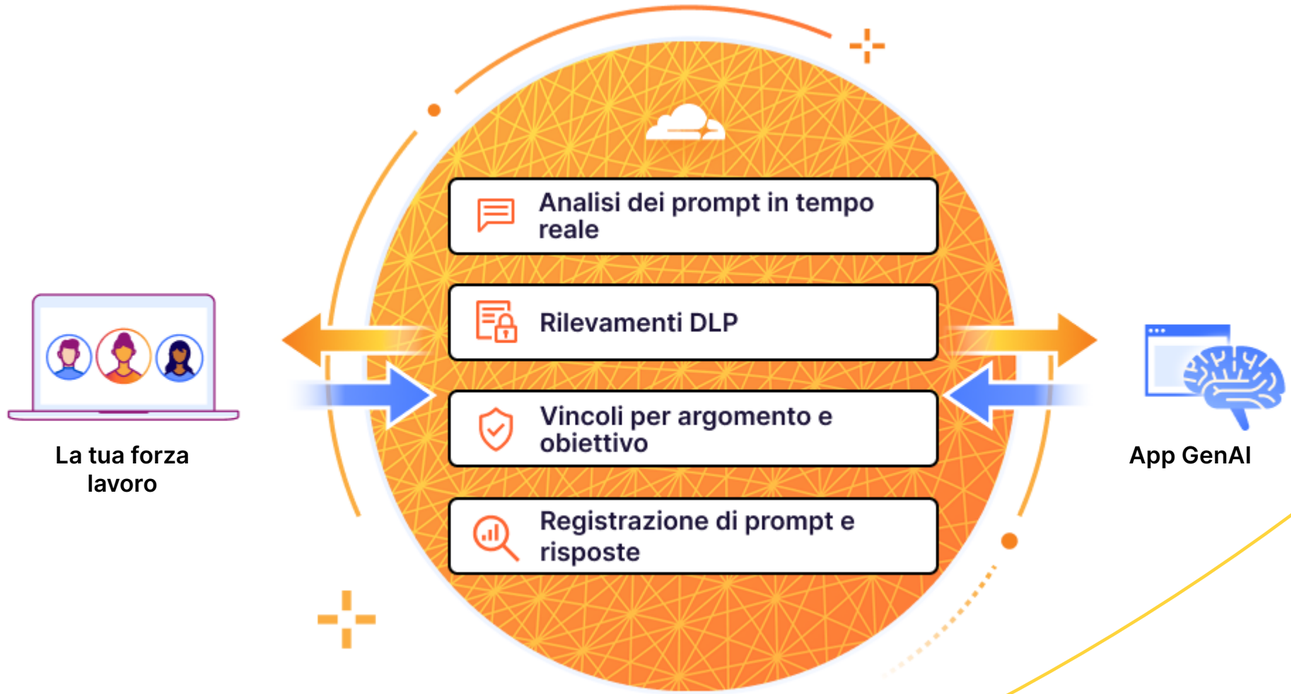
Il filtro DNS è efficiente per i suoi semplici criteri di "consenti o blocca". Ma per incoraggiare l'adozione dell'IA, le organizzazioni vogliono andare oltre questo approccio binario e, invece, concentrarsi sulla protezione dei dati quando gli utenti interagiscono con gli strumenti IA.

Con l'ispezione HTTP attivata, un SWG come Cloudflare può rilevare, bloccare e registrare qualsiasi tentativo dell'utente di includere dati sensibili in un prompt IA. In questo caso, Cloudflare utilizza i classici **rilevamenti di prevenzione della perdita di dati (DLP)** per informazioni di identificazione personale (PII), codice sorgente, dati dei clienti, informazioni finanziarie, credenziali e altro ancora.

L'SWG può analizzare non solo il ***contenuto,*** ma anche il ***contesto*** dei prompt IA per interrompere l'esposizione dei dati. In questo caso, ad esempio, Cloudflare cerca uno scopo inappropriato e dannoso in un prompt dell'utente e crea **vincoli in base all'argomento e all'obiettivo** per prevenire qualsiasi output rischioso. Quindi, se un prompt tenta di richiedere informazioni di identificazione personale (PII) o codice dannoso o cerca di aggirare i criteri di un modello IA, Cloudflare bloccherà e registrerà anche quello.

La realtà è che la maggior parte delle persone condividerà eccessivamente i dati con l'intelligenza artificiale. In effetti, un recente sondaggio ha rilevato che il **93% dei dipendenti** ammette di inserire informazioni negli strumenti IA senza approvazione.[5] I rilevamenti e i vincoli DLP aiutano i team di sicurezza a trovare un equilibrio tra l'incoraggiamento della produttività e la mitigazione dei rischi.

Protezioni e vincoli dei prompt in tempo reale con l'SWG di Cloudflare



5 Proteggi lo sviluppo dell'IA

Proteggi gli sviluppatori che creano app basate sull'intelligenza artificiale

Sempre più organizzazioni non solo stanno adottando strumenti IA, ma stanno anche creando internamente le proprie app basate sull'intelligenza artificiale. La distribuzione del filtro DNS protegge i team di sviluppatori responsabili della creazione di queste esperienze di intelligenza artificiale nei loro flussi di lavoro quotidiani. Lo stesso approccio collaudato può mitigare nuovi rischi, inclusi quelli seguenti:

- **Blocco dei tentativi di "model phishing" e di data poisoning:** le app IA fanno molto affidamento su librerie esterne, modelli pre-addestrati e dati da hub come Hugging Face e integrazioni API. Gli autori di attacchi possono eseguire il typosquatting domini che ospitano servizi di intelligenza artificiale contraffatti (ad esempio, il "quasi identico" *huggngface.co* invece del corretto *huggingface.co*). Gli sviluppatori possono raggiungere inavvertitamente queste destinazioni false digitando erroneamente o facendo clic su un collegamento. Lì, possono essere indotti con l'inganno a inserire credenziali API, scaricare codice dannoso o utilizzare modelli e set di dati avvelenati. **Un filtro DNS è in grado di intercettare e bloccare le query a questi domini rischiosi e spesso di nuova registrazione, prevenendo questi attacchi di phishing e della supply chain.**
- **Arresta l'esfiltrazione del peso del modello:** i pesi di un modello IA sono i suoi gioielli della corona, che rappresentano una proprietà intellettuale di alto valore. Una tattica di esfiltrazione comune consiste nell'utilizzare una macchina di sviluppo compromessa per caricare questi file su domini di condivisione file o repository privati oscuri. **I giusti criteri di filtro DNS (ad esempio, limitando la risoluzione DNS solo alle risorse autorizzate) sono in grado di bloccare la richiesta della macchina prima ancora che inizi qualsiasi trasferimento di dati.**
- **Blocca le inoculazioni di prompt indirette:** uno sviluppatore può incaricare un agente IA di analizzare una pagina web o un contenuto con istruzioni nascoste con obiettivi dannosi. Tali istruzioni potrebbero indicare all'IA di recuperare più dati da un dominio specifico o di eseguire un callback C2 a un server compromesso. **I filtri DNS possono impedire questa inoculazione di prompt indiretti impedendo il pulldown dei dati dell'agente o i tentativi telefonici di casa.**

Spotlight sulla piattaforma per sviluppatori di Cloudflare

Crea esperienze di intelligenza artificiale sicure e all'avanguardia

La piattaforma per sviluppatori di Cloudflare fornisce l'infrastruttura per scalare le tue applicazioni IA in ogni fase: crea app e agenti IA, archivia i dati di addestramento, esegui l'inferenza IA, con sicurezza fin dalla progettazione.

- **Controlla e osserva le app basate sull'intelligenza artificiale**, riducendo al contempo i costi di inferenza e instradando dinamicamente il traffico
- **Crea server MCP (Model Context Protocol)** con autenticazione e autorizzazione integrate
- **Previeni le interruzioni dei servizi** con fallback dei modelli e limitazione della frequenza

Uno sguardo al futuro: garantire l'adozione dell'IA con Cloudflare One

Inizia con il filtro DNS

Come soluzione autonoma, il filtro DNS offre un modo semplice ed efficace per affrontare le principali sfide e opportunità dell'intelligenza artificiale. Le distribuzioni con o senza un dispositivo client e la gestione intuitiva dei criteri aiutano i team di sicurezza e IT a iniziare a realizzare rapidamente valore attraverso la forza lavoro ibrida.

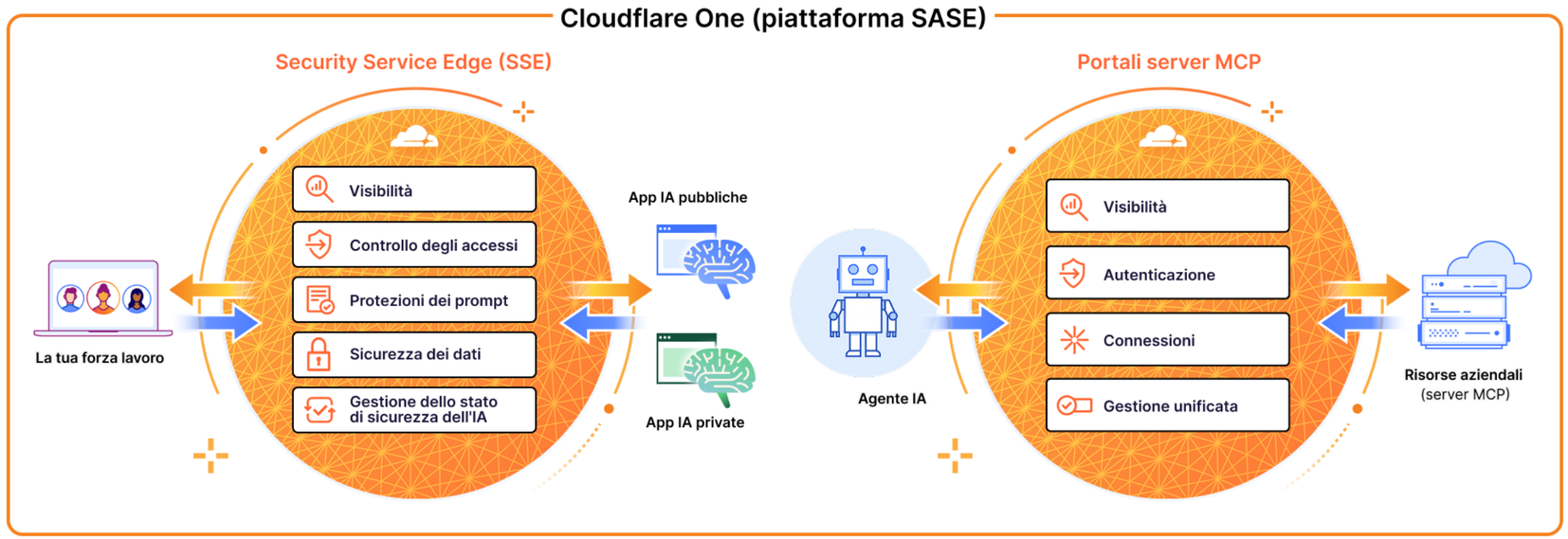
Per molte organizzazioni, la modernizzazione del filtro DNS è un primo passo comune verso una distribuzione SWG completa o un'architettura SASE consolidata. Piattaforme come Cloudflare che semplificano questa progressione consentono alle tue organizzazioni di adattarsi con agilità e accelerare in modo sicuro l'adozione dell'IA.

Oltre il filtro DNS

Estendi la piattaforma SWG e SASE per proteggere l'uso da parte della forza lavoro dell'IA generativa e agentica

Cloudflare One, una piattaforma SASE, si trova tra la tua forza lavoro e gli strumenti IA, rendendola un punto di controllo logico per proteggere l'uso degli strumenti IA. Sia che i dipendenti chattino con ChatGPT o che gli agenti IA raccolgano informazioni attraverso le risorse aziendali, Cloudflare offre visibilità e sicurezza coerenti sia nelle interazioni da uomo a intelligenza artificiale che da macchina a macchina, il tutto da un dashboard e un piano di controllo unificati:

- **Scopri la shadow AI** e gestisci i criteri per tutti gli strumenti di intelligenza artificiale autorizzati e non
- **Rafforza la governance dell'IA** con controlli degli accessi basati sull'identità per l'uso della GenAI e la comunicazione dell'IA agentica
- **Arresta la perdita di dati** bloccando le informazioni sensibili nei prompt utente, applicando misure di sicurezza specifiche ed eseguendo la scansione per individuare configurazioni errate nell'IA



Riferimenti

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. Ricerca di ManageEngine del 2025: 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 segnalata AI-orchestrated cyber espionage campaign”, Anthropic, 13 novembre 2025: https://www.anthropic.com/news/disrupting-AI-espionage
2. Ricerca di ManageEngine del 2025: https://www.manageengine.com/survey/shadow-ai-surge-enterprises/

+44 20 3514 6970 | enterprise@cloudflare.com | www.cloudflare.com/it-it/

REV:BDES-8725.2026JAN28