CLOUDFLARE

电子书

构建现代化安全体系以迎接 AI 时代

保护人员、数据和应用安全的全新实践指南

目录

内容摘要

如果可以重建您的安全架构，鉴于了解到 AI 所带来的影响，您还会采用类似现在的架构吗？

当我与全球各地的安全负责人交流时，得到的答案几乎都是“不”。

为什么？这是因为，当前 AI 落地大幅加速的时代背景，正迫使所有人重新评估自己的安全技术栈。即使在生成式 AI 成为主流之前，IT 和安全团队就已经需要艰难地管理数十个相互独立的单点解决方案。叠加更多工具只会进一步增加复杂性、成本和风险。

当 AI 本身成为新的攻击面时，以往将专业化、一流的最佳工具拼凑起来的做法变得不堪一击。

员工可能会以各种不可预测的方式使用和滥用 AI 和生成式 AI 应用。AI 和机器学习 (ML) 模型、框架、应用、智能体以及合规标准演变过快，孤立的工具难以跟上其步伐。而且，如今攻击者正利用 AI 加快攻击节奏，被动式的安全防护方案不足以应对，造成极高风险。

在这一新现实下，重新思考安全模型不仅是审慎之举，更是保持竞争力与韧性的关键。企业需要一种全新架构：既能助力驾驭 AI 的力量，又能守护自身系统安全，抵御各类新兴威胁。

下文介绍在当今时代取得成功所需的全新安全模型，以及助力自信驾驭 AI、并领先 AI 驱动威胁的策略。

Amit Chaudhry
解决方案营销高级总监，Cloudflare

为何 AI 需要新的安全模型

AI 正以前所未有的速度重塑商业格局，而能跟上的安全架构寥寥无几。新的模型、框架、协议（如模型上下文协议 (MCP)）、硬件进步和集成标准不断出现。诸如 Claude、ChatGPT 和 Copilot 等工具，赋能工程团队之外的业务人员，往往完全绕过 IT 部门审批流程。一些组织也开始使用 AI 进行构建，并推出由 AI 驱动的全新功能。

企业经常发现自己难以跟上最新技术发展的步伐，如果不能迅速适应，就有可能面临被市场淘汰、合规失效、声誉受损和陷入竞争劣势的风险。

AI 落地与创造，要求企业构建基于三大支柱的全新安全措施，为安全的持续创新奠定坚实基础：

保护人员和数据

保护员工并确保他们以安全的方式使用数据、设备、自动化服务、服务器和其他消耗生成式 AI 资源的系统。

保护 AI 驱动的应用

在整个 AI 应用开发生命周期中嵌入安全性，并保护应用本身，使其免受数据风险、模型操纵和其他恶意活动的影响。

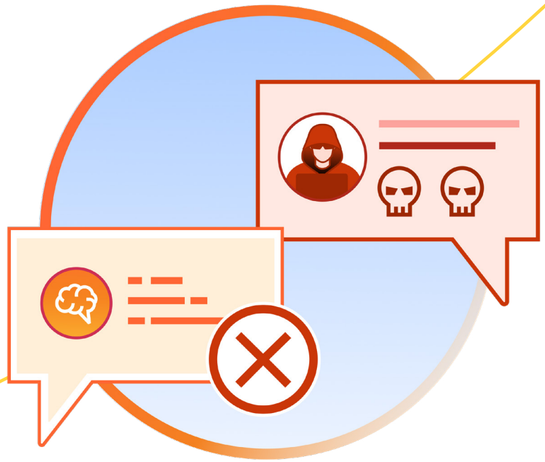
使用 AI 防御威胁

借助 AI 强化企业的安全态势并降低复杂性，提供高级威胁检测、自动化响应等能力。

本实践指南分享应对这些关键安全要务的关键策略。此外，本指南介绍组织如何通过将各种工具和基础设施整合到一个现代安全平台中来简化其架构。

AI 带来了一系列广泛的安全风险

对员工使用情况的可见性有限	**85%**	的 IT 负责人表示，员工在未经评估的情况下就已开始采用 AI 工具。[1]
不受控制的数据泄露和合规风险	**90%**	的员工信任未经授权的 AI 工具来保护其数据；50%的员工认为使用未经批准的工具几乎没有任何风险。[1]
不一致的治理	**74%**	的组织存在AI开发工具链碎片化问题，导致治理难度增加。[2]
不安全的应用开发	**37%**	的组织建立了部署 AI 工具前进行安全性评估的流程。[3]
对抗性 AI 突破传统防御措施	**80%**	的勒索软件攻击使用 AI，包括深度伪造和 AI 生成的网络钓鱼活动。[4]



保护使用 AI 时的人员和数据安全

如今，AI 领域增长最快的风险并非来自恶意黑客，而是员工在无意间将敏感数据上传至生成式 AI 应用。

例如，假设一名会计人员指示一个生成式 AI 工具搜索并显示公司当前月份的可计费工时。他们指示该工具在客户付款日期临近时发送发票和付款提醒。现在，AI 发起一笔需要追踪、记录，尤其是保持私密的交易。

员工远在正式政策出台之前就已经采用 AI，而且往往未意识到潜在风险。影子 AI 部署绕过了传统审查，造成隐蔽的攻击面并引入新的合规风险。

为 AI 应用制定使用规范是保障 AI 应用安全的关键第一步。**然而，真正的执行需要将访问控制和数据限制直接构建于 AI 交互界面之中。**

保障员工使用 AI 安全的五个必要措施

1

发现影子 AI 使用

组织无法保护他们看不见的东西，包括影子 AI 的使用。识别并过滤所有发送到互联网的 AI 流量；一旦发现使用生成式 AI 应用的情况，立即执行适当的策略。

2

检测和控制 AI 应用访问

应用 Zero Trust 安全最低权限原则，确保只允许授权的 AI 服务和受信任设备上的授权用户连接您的网络基础设施。

3

保护敏感数据

数据丢失防护 (DLP) 功能会阻止尝试通过有风险的生成式 AI 工具共享或上传专有代码、个人可识别信息 (PII) 和其他敏感数据的行为。

4

阻止有害或有毒的提示词

员工可能会不小心或有意地向 AI 服务提交不当的提示词或主题。阻止有害的交互，防范模型投毒和不正确的输出，并执行公司政策。

5

增强态势管理

部署具备云访问安全代理 (CASB) 功能的 AI-SPM 服务。CASB 可扫描生成式 AI 服务的配置错误及数据泄露风险。

> “尽管 AI 拥有大量合法用户，但其确实带来了重大安全和隐私问题。Cloudflare 帮助我们找到存在的影子 AI 风险，并阻止未经批准的 AI 应用和聊天机器人。”

Matthew Ortiz
信息安全高级经理， **Indeed**

了解详情 >

保护 AI 驱动的应用和工作负载

无论组织是基于内部数据自主构建大语言模型 (LLM)，还是在面向公众的应用（例如通过网站）中集成第三方 AI 工具，均存在重大风险。

例如，客户支持机器人如果被操控，可能会泄露敏感的员工数据或商业机密。攻击者也可能滥用模型，通过发送大量请求使模型过载，导致 AI 资源过度消耗或拒绝服务。诸如此类的 AI 特定威胁手段会将有用的 AI 工具变为负担。

换言之：来自自主开发和第三方 AI 应用的风险都在快速增长。

开放式 Web 应用安全项目 (OWASP) 发布了 LLM 十大风险报告，警示企业组织注意敏感信息泄露、提示词注入、模型投毒以及适用于任何 AI 工具的其他 LLM 威胁。

尽管存在这些警示，但仅有 37% 的组织已制定在部署前对 AI 工具安全性进行评估的流程。[3]

应用部署之后才添加安全防护措施或尝试从外部模型中追回数据，为时已晚。

有效的安全防护必须考虑模型层本身——涵盖组织构建、采购及使用的所有 AI 工具的生态系统。

保护 AI 应用的三项要求

1

扩展对应用和 API 的可见性

专业 AI 防火墙能够发现并标记生成式 AI 和应用程序编程接口 (API) 端点，检测泄露 PII 的行为，并阻止恶意提示词影响 AI 模型性能，防止恶意内容或虚假信息导致模型中毒。

2

在边缘阻止威胁

实时威胁检测和缓解功能全面阻止 AI 管道中的 AI 特定漏洞、错误配置和攻击手段，例如提示词注入、数据投毒和模型滥用。

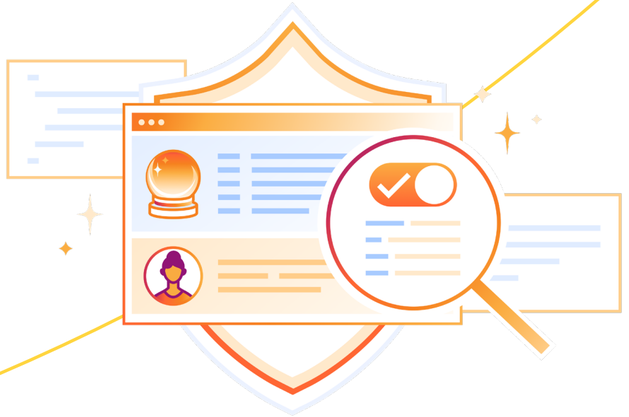
3

确保合规开发

AI 感知数据保护主动管理数据输入，在 AI 模型和管道中执行严格的访问控制，维护审计跟踪，并帮助遵守欧盟 AI 法案等相关要求。

> **“随着我们进一步转向基础设施即代码、可编程性与自动化、AI 并更深入理解如何通过 Zero Trust 提升云安全与可见性，Cloudflare 的低成本优势与全球覆盖为我们提供了无限的机遇。”**
>
> **Michael Lee**
> **网络工程经理，VistaPrint**
>
> 了解详情 >

vistaprint.



使用 AI 防御威胁

由于 AI 擅长模式识别，攻击者现在可以更轻松地识别和利用传统防御措施中的弱点。例如，威胁行为者正在利用 AI：

- 制作极具说服力的网络钓鱼和社会工程学诈骗，绕过传统安全措施
- 部署自动化机器人，以安全团队难以及时响应的规模和速度利用网络漏洞
- 开发新型手段，包括更复杂的身份欺诈方式

随着攻击者继续利用 AI 作恶，从被动转变为主动防御、AI 驱动的安全防护不再是可选项 ——而是必然之举。

为了应对这些及其他威胁，组织的安全架构必须不断调整和演进，才能保持领先。

AI 驱动的防御能够适应每个组织的独特行为和需求。这种定制化方法能够在整个组织内实现更精准的威胁检测、更快速的自动化事件响应和预测性的风险管理。

构建主动式 AI 驱动安全体系的五项策略

1

利用 AI 驱动全面的安全防护和可观测性能力

在单一界面中统一日志、分析、警报和取证，以识别风险及其根本原因。

2

利用 AI 驱动的安全防护，实时检测并消除威胁

利用 AI/ML 赋能的检测来分析庞大的数据集、识别异常并自动响应新兴威胁。

3

防御 AI 驱动的钓鱼攻击、深度伪造和恶意软件

部署基于 AI 的检测、防钓鱼身份验证和自适应安全控制以应对不断演变的攻击。

4

保护专有数据，防范 AI 抓取和自动化威胁

利机器人管理、API 身份验证和数字水印服务来防止数据盗窃和利用。

5

利用 AI 驱动的行为分析，实时发现内部威胁

持续分析用户行为（而不仅仅是身份），以评估异常访问模式、权限提升和数据窃取尝试。

> “GPT4 发布后，我们分析了 Cloudflare 平台的统计数据，发现一个月内攻击增加了 20%。我们依赖 Cloudflare 的电子邮件安全解决方案，该方案能够很好地处理日益增长的网络钓鱼攻击，并帮助我们保护用户和公司。”

Roman Bugaev
首席技术官，**Flo Health**

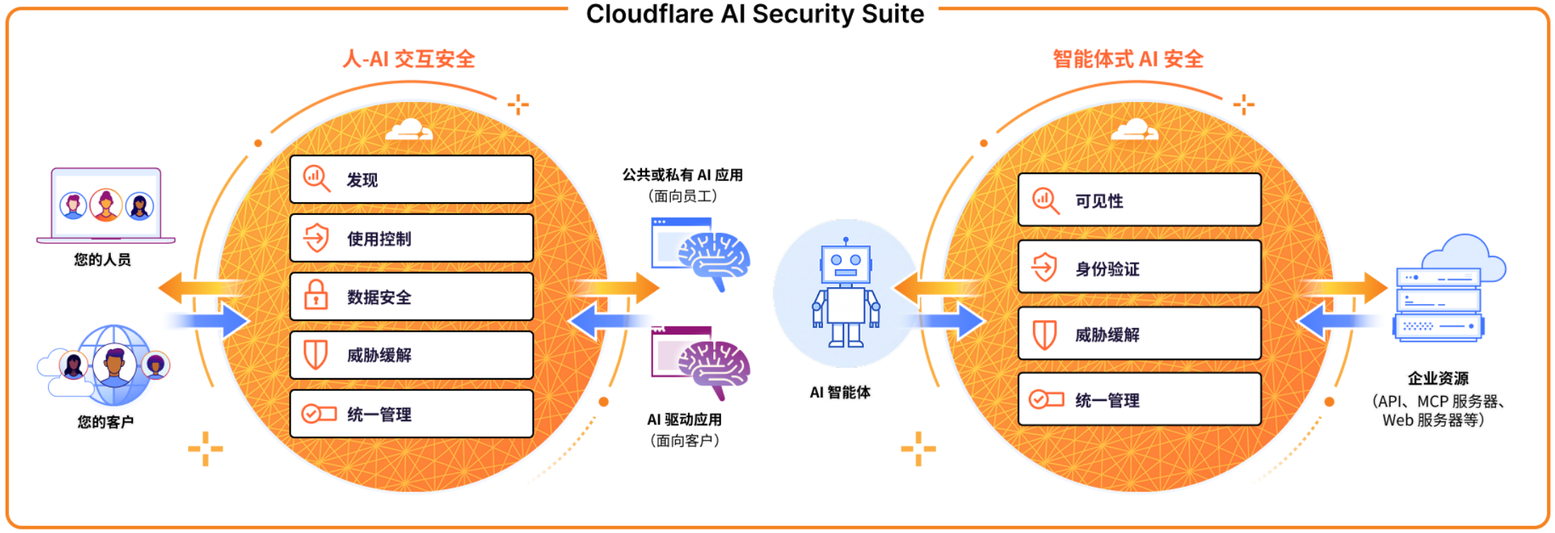
了解详情 >

保护人员、数据和应用的一体化解决方案

要在不影响安全的前提下加速 AI 应用落地，组织需要一个由可编程的全球网络驱动，整合安全、连接和开发服务的一体化平台。

Cloudflare 全球连通云为提供安全、低延迟和无限可扩展的连接，覆盖应用、全球用户和客户群、API 和混合网络。这些优势有助于组织简化复杂性、增强安全性和合规性，并缩短 AI 转型的价值实现周期。

我们的 AI 安全服务以 Cloudflare 的全球连通云为基础构建，共同协作以提升 AI 工作负载的安全性、可扩展性和效率。这些服务构成单一的端到端解决方案，旨在保护人员、数据和应用免受新兴威胁，同时优化性能。



1. Cloudflare 如何保障使用 AI 时的人员和数据安全

安全访问服务边缘 (SASE) 本已成为整合安全和网络功能的关键组件，如今更是实现有效 AI 安全的核心基础。

Cloudflare 的 SASE 套件——Cloudflare One 为 AI 使用提供多方面的安全防护。企业可获得实时威胁检测、治理与防护工具，从而确保数据安全、员工高效工作以及业务运营的韧性。

发现：

- **影子 AI 报告**揭示员工如何使用 AI 工具。
- **带外 API CASB 集成**检测错误配置，并提供对主流公共 AI 工具内部安全态势的可见性。

衡量：

- **AI 置信度评分**用于评估与第三方 SaaS 和 AI 应用相关风险，以便组织能够制定相应的策略。

保护：

- **AI 提示词保护**是一种 DLP 功能，可防御恶意输入并防止敏感数据丢失。

保障：

- **MCP 安全**集中策略实施，提供全面的可见性和日志记录，简化复杂的 MCP 部署。

2. Cloudflare 如何保护 AI 驱动的应用和工作负载

需要采取对开发人员友好的防护措施，以减轻在应用开发中纳入有害、不准确或不适当内容的风险。

Cloudflare 的 AI 解决方案 — AI Gateway、Firewall for AI 和 Workers AI — 协同工作，提供端到端解决方案，保护 AI 驱动的应用免受新兴威胁侵害。

AI Gateway 充当应用和 AI 模型提供商之间的中央节点（代理），以便：

- **实施** AI 防护措施，以防止有害的提示词和 AI 响应
- **应用**高级速率限制以防止模型滥用
- **安全**存储对接模型提供商（如 OpenAI、Anthropic 等）的 API 密钥

Firewall for AI 提供为 LLM 漏洞量身定制的防御层，以便：

- 使用基于 AI 威胁情报训练的检测引擎**分析**传入的提示词
- **识别**并阻止恶意模式、异常和违反策略的行为
- **确保**只有安全的提示词能够到达模型

Workers AI（AI 推理即服务）确保 AI 推理本身在全球分布式的高性能安全环境中运行。

Workers AI 不使用用户提示词来训练模型，并托管 50 多个开源模型，确保透明度。

组织可以获得智能体和 AI 应用所需的 AI 推理工作负载，并可开发具有 **内置集成 (OAuth 2.1) 身份验证和授权功能的 MCP 服务器。**

3. Cloudflare 如何利用 AI 进行防御

Cloudflare 的智能、自适应安全态势用 AI 防御对抗性 AI。

每天，Cloudflare 庞大的全球网络分析互联网上数万亿个互联网信号。这包括每秒处理 8400 万个请求和 6100 万个 DNS 查询等。这种数量、速度和多样性均无可比拟的流量提供了广泛的威胁遥测数据。因此，我们的机器学习模型能够为组织的全部数字资产提供先进、主动的防护。

Cloudflare 网络的强大力量支撑着我们防御复杂攻击的方法，包括针对 AI 系统的攻击，以及恶意利用 AI 的行为。

> **“……随着生成式 AI 的快速发展，我们预计行业将面临新一代的复杂网络威胁。我相信 Cloudflare 拥有所需的可见性、资源和和商业情报，能够利用全球网络上的海量数据并训练防御模型，有效识别和缓解这些下一代攻击。”**
>
> **Mehdi Salour**
> 全球网络与 DevOps 高级副总裁，**8x8**

多手段 AI

防范不断演变的网络威胁，包括利用 AI 技术增加攻击数量和真实性的威胁。

定制解决方案

分析特定于组织的流量模式，并调整策略以适应其独特的行为和环境需求。

高级 AI 与实时分析

增强应用、API、网络、人员以及 AI 工作负载本身的威胁检测、缓解和保护能力。

准备好构建现代化安全体系了吗?

Cloudflare 为企业、开发者、初创公司以及数字媒体和内容所有者提供所需的工具和基础设施，助力其自信地应对 AI 落地过程中的各种复杂挑战。

借助 Cloudflare 广泛的全球边缘网络（覆盖 125 个国家 / 地区的 330 多个城市）、集成式安全解决方案及以开发者为中心的工具，Cloudflare 助力赋能组织更快创新、高效自动化，并构建安全的 AI 驱动应用，同时保障其数据与应用安全。

进一步了解如何为 AI 创新奠定安全基础。

申请定制企业演示

延伸阅读和资源

为网络安全领域的 AI 未来做好准备

今天如何开始制定战略，帮助企业在未来实现 AI 的价值最大化？Cloudflare 首席安全官 Grant Bourzikas 分享了四项主要建议。

阅读文章

Cloudflare AI Security Suite

自信推进 AI 落地。了解 Cloudflare AI Security Suite 如何保护整个 AI 生命周期，以防御诸如提示词注入、数据泄露和模型滥用等攻击。

阅读解决方案简述

安全信号：对抗性 AI

Cloudflare 现场首席信息官 Khalid Kark、Accenture 董事总经理兼 AI 安全主管 Daniel Kendzior 以及 Cloudflare 首席信息官 Mike Hamilton 深入探讨了 AI 在网络安全领域的实际影响。观看视频，了解如何利用 AI 进行防御，而不仅仅是检测。

观看采访

使用 SASE 保护生成式 AI 的最佳实践

本指南深入介绍管理（人类）员工访问 AI 的三个关键支柱：可见性、风险管理和数据保护，以及有关使用 MCP 在企业中部署智能体式 AI 的指导原则。

阅读博客文章

参考资料

1. ManageEngine。“97%的 IT 决策者认为影子 AI 存在重大风险，而 91% 的员工认为影子 AI 没有风险、风险极低，或者回报超过风险。” 新闻稿，2025 年 7 月 8 日。https://www.manageengine.com/news/shadow-ai-report.html。
2. Anamarija Pogorelec。“弥合 AI 模型治理差距：为 CISO 提供的主要发现”。HelpNetSecurity，8 月 18 日2025 年。https://www.helpnetsecurity.com/2025/08/18/ciso-ai-model-governance/。
3. Sadie Creese 与 Akshay Joshi。“领导层指南：管控 AI 落地带来的网络风险。” 世界经济论坛，2025 年 1 月 21 日。https://www.weforum.org/stories/2025/01/a-leaders-guide-to-managing-cyber-risks-from-ai-adoption/。
4. Church, Zach. “现有 80% 的勒索软件攻击使用 AI。” 麻省理工斯隆商学院，2025 年 9 月 8 日。https://mitsloan.mit.edu/ideas-made-to-matter/80-ransomware-attacks-now-use-artificial-intelligence。

CLOUDFLARE

010 8524 1783 | enterprise@cloudflare.com | www.cloudflare.com/zh-cn/

REV:BDES-7856.2025OCT06