AI 安全：
您在 AI 竞赛中的竞争优势

企业高管指南：加速 AI 落地并管控风险

引言

随着企业急于利用 AI 的变革力量，AI 采用的速度往往快于安全防护建设。每一项经过正式审批的 AI 计划背后，都存在着无数个人或团队进行的影子 AI活动。事实上，85% 的 IT 决策者表示，在 IT 团队能够进行评估之前，员工们就已经开始采用 AI 工具了[1]。93% 的员工承认在未经批准的情况下将信息输入 AI 工具[1]。

与此同时，AI 原生攻击活动正在激增，过去一年中增长了 47%[2]，传统安全工具难以应对。新型数据暴露方式和合规缺口对现有治理实践构成挑战。企业高管正面临一个关键问题：应如何管控 AI 带来的风险，同时又不阻碍其催生的创新？

要求很明确。企业需要创建这样的环境：

- 所有 AI 使用了如指掌
- 所有 AI 通信得到保护
- 所有 AI 策略均可执行
- 所有 AI 模型免受滥用

Cloudflare AI Security Suite 消除风险的不确定性，让组织更有信心地加速推进 AI 落地。我们的统一平台发现 AI 使用情况，应用 Zero Trust 访问，并通过主动威胁防御保护 Web 和 API 端点，有效保障企业的整个 AI 生命周期安全。借助集成式数据治理，企业团队可自由创新，且无需牺牲安全性。

85%
的 IT 决策者表示，在 IT 团队能够评估之前，员工就已经开始使用 AI 工具[1]。

AI 安全挑战

AI 采用风险正在推动 AI 安全解决方案市场快速增长。例如，云原生应用保护平台 (CNAPP) 专注于 AI 开发工作流程。此类早期产品是一道重要的防线，但它们并不足够。企业需要保护整个 AI 生命周期，包括保护在生产环境中运行的 AI 系统。

AI 安全挑战的完整范畴，正变得愈发清晰。如今，开发人员正在构建 AI 功能，员工正在使用外部 AI 工具，客户正在与 AI 驱动的应用进行交互。手动保护所有这些不同的环境十分复杂，而要实现一致保护则难上加难。

防止员工不当使用 AI

随着组织员工同时使用经批准和未经批准的 AI 工具，敏感数据和运营面临风险。安全团队必须处理以下问题：

- **缺乏对 AI 使用情况的可见性：** 管理层通常无法全面掌握员工正在使用什么 AI 工具，敏感数据在何处被处理，以及这些 AI 系统如何与业务应用相连接。这一盲点造成了重大风险暴露。
- **数据安全与合规风险：** AI 正在改变数据在企业内部的流转方式。个人信息、专有数据和客户记录最终进入 AI 系统时，可能会触发合规违规行为或暴露商业竞争情报。
- **智能体式 AI 工作流的访问控制：** 不仅需要管理人员对 AI 工具的访问。智能体式 AI 对 MCP 服务器和其他关键系统的访问也必须进行管理。这需要采用新的身份管理和访问控制方法。

保护面向公众的 AI 应用和模型

无论是内部开发还是通过第三方使用，AI 系统正在成为客户和用户体验的核心支柱，因此需要提供相应的保护。OWASP 的 LLM 应用十大安全风险强调了传统工具无法解决的威胁：

- **无限制消耗攻击：** 与传统 DoS 攻击类似，无限制消耗攻击试图通过资源密集型请求压垮 LLM 。按使用量付费的云定价模式会导致此类攻击的财务影响飙升，同时合法用户还会面临服务质量下降的问题。
- **模型投毒：** 攻击者将损坏的数据注入开发人员用于模型训练的公共数据集或存储库，从而在 LLM 中植入后门、偏见或漏洞。以这种方式中毒的模型会表现正常，直到特定触发因素激活有害行为。
- **提示词注入攻击：** 提示词注入是一种通过 AI 界面窃取数据的常见手段。它通过在用户提示词或外部内容中嵌入恶意指令来操控 LLM 的输入，导致模型忽略原始指令，转而执行攻击者的命令。
- **越狱：** 精心编写的提示词（如角色扮演情景、指令覆盖命令和多轮策略）可以绕过 LLM 安全防护措施，以生成违禁或提取敏感信息。

为了实现 AI 的快速创新而不让企业面临风险，我们建议领导者采用一种涵盖全方位要求的 AI 安全方案：

- 确保员工安全使用生成式 AI
- 保护 AI 驱动的应用和工作负载
- 构建设计安全的 AI 驱动应用

保护 AI 开发和训练工作流

AI 项目涉及海量数据集、高成本资源及迭代式实验——这些均会产生新的攻击面与漏洞。安全团队必须解决以下问题：

- **训练数据的安全性与完整性：** 被篡改的训练数据可能引入偏差、后门或漏洞，并持续存在于生产模型中。企业必须确保对训练数据集的安全访问，防止未经授权的修改，并在确保整个模型生命周期的数据溯源。
- **凭据和密钥管理：** AI 开发工作流需要访问各种系统，包括数据集的云存储、用于训练的计算集群、模型仓库和第三方服务。机密或 API 密钥保护不力可能会暴露敏感凭据，导致对专有模型、训练数据或生产系统的未经授权访问。
- **开发环境访问控制：** AI 工程师通常需要更高权限，以便进行模型实验并访问敏感数据。如果没有适当的访问控制，这可能会导致内部威胁、意外数据泄露或未经授权的模型提取。

通过 Cloudflare 统一安全性

成功 AI 落地重点在于提升而非抑制生产力。当团队知道适当的防护措施已就位，从而充满信心地使用 AI 时，他们将能更快创新，并承接更具挑战性的项目。

Cloudflare AI Security Suite 为 AI 创新打造安全环境。通过统一安全访问服务边缘 (SASE) 和 Web 应用安全能力，企业高管能够连接并保护两个核心领域：

- 外部、面向公众的 AI 应用
- 内部、私有的 AI 系统和工作负载

Cloudflare AI Security Suite 专注于整个 AI 生命周期，它满足从发现、风险管理到数据保护的各种安全需求，同时保护用户访问以及 AI 驱动的应用和开发工作流程。为了提供关键的实时生产安全层，Cloudflare 全球网络以内联方式检查并过滤每一个 AI 交互，从而保护所有用户和应用的数据。

Cloudflare 并非仅在问题发生后检测，而是在问题和威胁到达 AI 模型之前就进行预防和阻止。安全团队获得领先于新兴威胁所需的可见性。

Cloudflare 的核心功能

Cloudflare AI Security Suite 将全面监控、实时防护及主动风险管理集于一身，提供全方位的 AI 安全解决方案。

全面的 AI 发现与可见性

有效的 AI 安全依赖于对经批准和未经批准的 AI 资源及使用情况进行完整、实时的盘点。Cloudflare AI Security Suite 的基础是持续监控和自动发现，以识别所有各类环境（公共、私有或内部）中的 AI 模型、助手、智能体和影子 AI 部署。

积极主动的 AI 风险管理

Cloudflare AI Security Suite 检测和缓解 AI 特有的漏洞、错误配置和攻击路径，从而帮助预防攻击，包括 OWASP LLM 十大风险。应用置信度评分有助于确定修复的优先级，以便团队首先处理最重要的风险。

AI 驱动应用的安全防护

为了帮助 SecOps 及时了解最新的 AI 威胁手段，Cloudflare AI Security Suite 提供针对 AI 特有漏洞、错误配置和 AI 管道内部攻击路径的主动威胁检测与缓解，包括防范提示词注入、数据投毒和模型滥用。

专用 AI 防火墙能够发现并标记生成式或智能体式 AI 以及 API 端点，检测泄露个人可识别信息 (PII) 的企图，并阻止恶意提示词，以避免其影响 AI 模型性能或通过有害内容或错误信息污染模型。

适用于生成式 AI 和智能体式 AI 工作流的 Zero trust 访问

最低权限之类的 Zero Trust 原则同时适用于员工队伍和 AI 智能体。Cloudflare AI Security Suite 可以对人与 AI 之间以及 AI 与 AI 之间的交互实施 Zero Trust 网络访问 (ZTNA) 策略。适用于 MCP 服务器的集中式日志记录和控制，确保智能体式 AI 仅访问其已获授权访问的内容，包括即时工作流程。

AI 感知型数据保护

有效的 AI 安全需要具备数据丢失防护能力，这种能力利用多语言模型来理解提示词的内容及其背后的意图。Cloudflare AI Security Suite 将数据丢失防护 (DLP) 功能融入训练、提示词和响应中，以防止 AI 模型和管道中的 PII 暴露、数据泄露和未经授权访问。以内联方式部署，以 API 为中心的运行时安全充当快速、简单的第一层防御，旨在补充 CNAPP 支持的左移策略。

数据本地化

LLM 和 AI 应用与其他数据环境一样，受到相同的法规约束。Cloudflare AI Security Suite 帮助企业确保 AI 工作负载符合地理和管辖范围限制，并通过相关策略，将训练数据和推理请求控制在已批准的区域内。

AI 开发的安全设计

与所有类型的软件一样，AI 应用的安全性也应在开发周期之初即内置其中，而非在后期额外添加。Cloudflare AI Security Suite 为开发者提供工具和框架，以构建设计安全的 AI 驱动应用。

Cloudflare AI Security Suite 的业务影响

AI 的崛起不仅仅是一次演进，而是一场堪比工业化或计算机化的颠覆性变革。因此，快速而有效地采用 AI 不仅关系到推动增长，更关系到企业的生存能力。推进缓慢或以不安全的方式落地，如今已成为各行各业关乎生存的重大威胁。

Cloudflare AI Security Suite 支持安全、可控且高效的转型，从而在竞争激烈的环境中满足不断提升的客户期望与市场需求。

- **加快 AI 创新：**当安全措施能够确保安全使用，而非禁止采用时，员工和团队可以探索新的 AI 应用，从而提高日常工作效率。完善的安全框架支持开发人员信心十足地构建强大的 AI 功能，而不会不危害敏感数据或系统安全。
- **降低 AI 相关风险：**企业可以管控 AI 采用带来的所有风险，包括 AI 驱动型 Web 应用中的固有 AI 相关风险。对开发中的 AI 应用采取安全措施，以确保员工不会泄露敏感数据或在 AI 训练集中暴露这些数据。SecOps 能够主动识别并缓解 AI 特有威胁和漏洞，从而最大限度地减少受攻击面，并保护关键数据和模型。
- **优化安全运营：**集中化可见性和对 AI 安全态势的控制，有效简化管理并优化事件响应效率。SecOps 团队可以专注于战略计划，不必疲于应对 AI 相关事件。
- **强健的数据治理与合规性：**AI 专用数据保护控制可帮助保护敏感信息，并满足整个 AI 生命周期中不断变化的法规要求。
- **降低总拥有成本 (TCO)：**相比于为每项 AI 安全挑战实施独立单点解决方案，利用集成现有安全投资的整合平台更具经济效益。

Cloudflare AI Security Suite 应用场景

Cloudflare AI Security Suite 非常适合解决 AI 落地的相关核心需求。

- **保障员工 AI 工具使用安全：**为员工对公共生成式 AI 工具（如 ChatGPT）和内部开发 AI 驱动应用的访问实施 Zero Trust 策略。
- **保护面向公众的 AI 应用：**保护集成 AI 模型的 Web 应用和 API（如聊天机器人和推荐引擎），防范可能暴露敏感数据或滥用模型的攻击。
- **影子 AI 管理：**自动发现企业内部未经批准的 AI 工具，并施加适当管控，以在可控风险范围内支持持续创新。
- **针对 AI 交互的 AI 驱动 DLP：**防止敏感数据在 AI 提示词或响应中暴露，确保 PII 和机密信息受到保护。
- **AI 开发安全：**为工程团队提供框架，使安全开发成为默认方式，支持其在不影响安全性的情况下快速构建 AI 功能。

实施注意事项

作为一项基础安全策略，应审慎对待 AI 安全。

从您现有的基础设施入手	最成功的 AI 安全实施是在现有 SASE 和应用安全工具的基础上构建，而不是取代它们。这种方式既能充分利用当前的投入，又能拓展防护范围以覆盖 AI 特定风险。
部署统一的内联保护	在网络边缘实时阻止恶意活动的能力对于 AI 安全至关重要。部署实时内联控制，并使用基于 API 的监控作为补充。
确保完全覆盖	您的 AI 安全策略应涵盖全方位的需求：员工使用生成式 AI 工具的安全，保护 AI 驱动的应用和工作流，保护智能体式 AI 工作流，以及 AI 开发工作流安全。
企业规模规划	选择能够随您的 AI 采用同步扩展规模的解决方案。适用于一个试点项目的方案，必须也能随着 AI 使用扩大而在扩展到整个企业范围。
验证您的成功标准	在实际场景中验证解决方案，然后再做出全面承诺。选择支持免费、自助激活其企业级功能的平台。这使您能够在小规模场景下测试完整的安全套件（例如针对单个团队或应用），从而快速验证其价值并确保其符合您的成功标准。

借助 Cloudflare AI Security Suite 迈出下一步

随着全球 AI 应用加速普及，能够安全地实现 AI 规模化部署的组织，将拥有显著的竞争优势。关键在于认识到，AI 安全并非要阻止 AI 的使用，而是合理明智地推动 AI 落地。

Cloudflare AI Security Suite 支持企业自信开展创新。Cloudflare AI Security Suite 在我们广受采用的 SASE 及应用安全平台之上构建，并对其功能进行了拓展，提供一体化 AI 发现、Zero Trust 访问控制、情报驱动的主动威胁防御，以及稳健的数据治理能力。通过全方位保护 AI 使用和模型，企业能够支持开发人员更快构建，赋能员工更高效工作，同时不会影响最终用户体验。

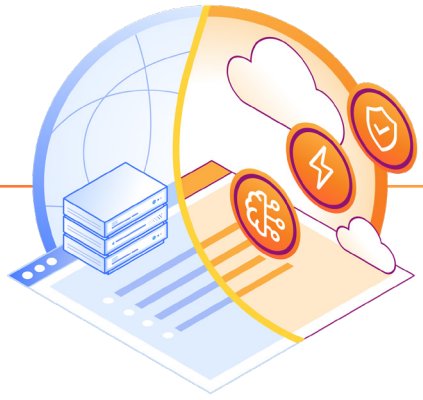
预约咨询

探索 Cloudflare AI Security Suite 如何转变企业安全推进 AI 落地的方式。

010 8524 1783

enterprise@cloudflare.com

www.cloudflare.com



1. ManageEngine。企业影子 AI 激增：来自美国和加拿大的洞察
2. Check Point Research。Check Point 软件公司发布的 2025 年第一季度全球网络攻击报告显示：全球网络威胁激增近 50%，勒索软件攻击增长 126%