CLOUDFLARE

电子书

AI 实践安全指南

CISO 指南：如何创建可扩展的 AI 战略

目录

内容摘要

欢迎，CISO!

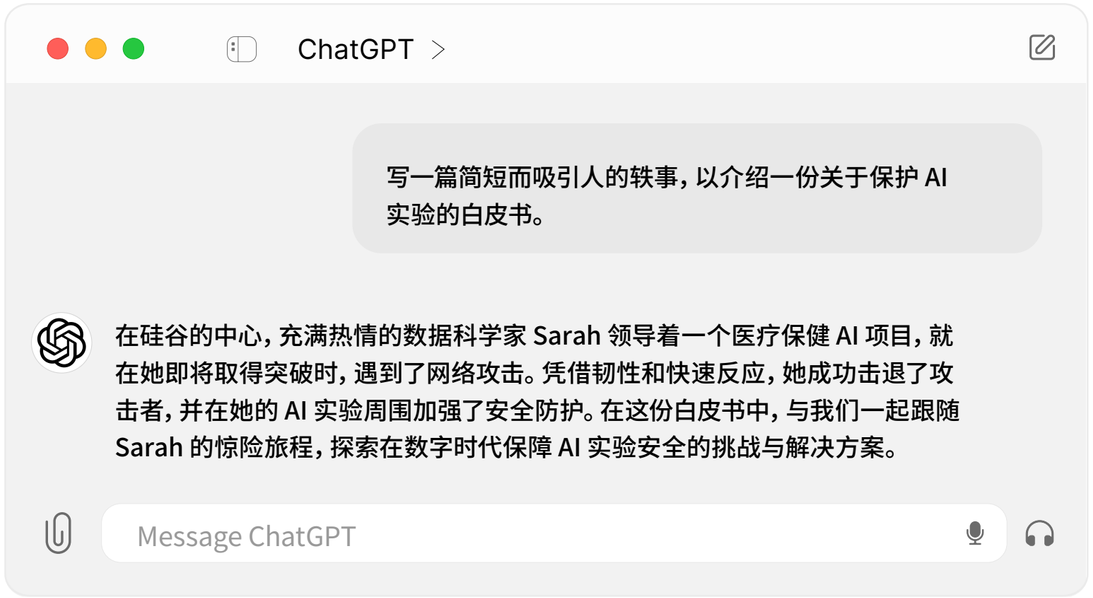
人工智能 (AI) 可能是最近最热门的一个词，也是安全社区最紧迫的问题之一。其影响力需要我们关注，因此 Cloudflare 编写了这份指南，旨在帮助您思考如何在组织内安全地进行生成式人工智能 (GenAI) 实验。

AI 工具正在迅速变得更强大、更易获取，为各行各业的创新释放了新的机遇。然而，与其他范式转变一样，GenAI 也带来了独特的安全、隐私和合规性挑战。GenAI 的广泛采用可能会引发意想不到的使用激增、用户滥用、恶意行为和危险的影子 IT 实践，这些都会增加数据泄露和敏感信息泄漏的风险。

随着 GenAI 在企业工作场所中的应用扩大，您需要准备一份 GenAI 蓝图，提供有关大规模使用、构建和保护的指南。让我们来讨论一下风险，并回顾您的团队可以根据成熟度和使用来保障 GenAI 安全的建议。利用这些策略，贵组织可以制定一个适合业务需求的 GenAI 战略，同时保护数据并确保合规。

- Dawn Parzych，产品营销总监，Cloudflare

保护生成式 AI 实验



抱歉，Sarah 的故事到此为止。虽然我们要和这个虚构人物说再见，但随着预测性和生成式 AI的不断扩展，现实生活中会出现无数个 “Sarah” ——她们是 IT 和开发团队中的主角，业务技术专家和普通员工。

AI 让技术专家和普通用户都为之着迷，激发了他们的好奇心和探索欲。随着我们努力释放 AI 的全部潜力，这种实验都是必要的。但是，如果缺乏谨慎和防护措施，它也可能导致安全或合规问题。

为了达到平衡，并更有效地理解和管理 AI 发展倡议，组织必须考虑三个关键方面：

1 利用 AI

使用第三方供应商提供的 AI 技术（例如 ChatGPT、Bard 和 GitHub Copilot），同时保护资产（例如敏感数据、知识产权、源代码等），并根据用例缓解潜在风险

2 构建 AI

开发针对组织特定需求定制的 AI 解决方案（例如用于预测分析的专有算法、面向客户的助手或聊天机器人，以及 AI 驱动的威胁检测系统）

3 保护 AI

保护 AI 应用和 AI 系统，防止恶意行为者操纵，使其出现不可预测的行为

保护生成式 AI 实验

生成式 AI 转型：今天与未来

生成式 AI 对消费者和组织的吸引力使其采用呈现前所未见的发展势头。一小部分高级用户迅速增长，在一定程度上要归功于活跃的开源社区，以及消费者驱动的应用（例如 ChatGPT 和 Stable Diffusion）等实验。

这个过程中，用户们发现，机器人实际上并不会“取代我们”。

生成式 AI 将人类置于不断完善和增强的位置，而不是从头开始创造一切，并可以帮助企业提升人员工作效率。预测式 AI 也提供类似的优势，帮助更容易利用数据来改善决策、打造更智能的产品和个性化客户体验，等等。

如今，**59% 的开发人员**在他们的开发工作流程中使用 AI 技术[1]

到 2026年，超过 **80% 的企业**将在生产环境中使用基于生成式 AI 的 API、模型和/或应用（目前仅 5%）[2]

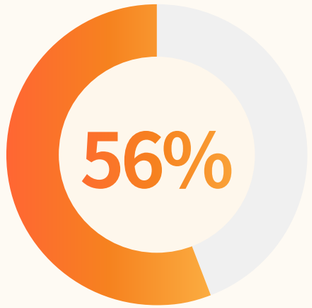
到 2030 年，生成式 AI 将增强**知识工作者 50%** 的任务，以提高生产力或改善平均工作质量（目前不到 1%）[3]

1. SlashData，“开发人员如何与人工智能技术互动”，2024 年 5 月
2. Gartner，“A CTO’s Guide to the Generative AI Technology Landscape”（首席技术官的生成式 AI 技术形势指南），2023 年 9 月
3. Gartner，“Emerging Tech: The Key Technology Approaches That Define Generative AI”（新兴技术：定义生成式 AI 的关键技术方法），2023 年 9 月

安全使用生成式 AI

AI 实验涵盖从使用预构建 AI 工具和服务到从头开始构建自定义 AI 解决方案的一系列领域。一些组织可能向创建自己的 AI 模型和应用发展，许多组织则会坚持利用第三方 AI 工具。

在这些情况下，第三方 AI 工具会带来新的风险，因为组织对其安全和隐私配置仅掌握有限的直接控制权。



56% 的员工在工作中使用了 AI 驱动的工具。但只有 26% 的公司制定了 AI 政策[4]

员工们可能正在通过像 Microsoft 365 这样的 SaaS 套件、内置于搜索引擎或公共应用中的聊天机器人、甚至 API来使用现成 AI 工具开展工作。

组织必须进行尽职调查以最大程度减少风险，包括：

- **评估**第三方工具的安全风险
- **解决**数据隐私问题
- **管理**对外部 API 的依赖（或过度依赖）
- **监控**潜在漏洞

一个例子是在员工使用公共 Web 应用（如 ChatGPT）时。输入到提示词中的内容都成为离开组织控制的数据。用户可能会过度共享敏感、机密或受监管信息，如个人可识别信息 (PII)、财务数据、知识产权和源代码。即使他们不共享明确的敏感信息，通过拼凑输入的上下文也可能推断出敏感数据。

作为防护措施，员工可切换一项设置来防止其输入被用于模型的进一步训练，但他们必须手动操作。为确保安全，组织需要找到方法来防止员工输入私有数据。

为 AI 的安全影响做好准备

数据暴露

用户与外部 AI 服务不当共享敏感数据的情况有多严重？匿名化/假名化技术是否足够？

API 风险

您将如何解决第三方 API 中可能成为攻击者潜在入口的漏洞？

黑箱系统

外部 AI 模型哪些决策过程可能引入意外风险？

供应商风险管理

您对第三方 AI 提供商的安全实践了解多少？更重要的是，什么是您不知道的？

4. The Conference Board，2023 年 9 月

保护 AI 使用的步骤

1 管理治理与风险

- 制定关于如何和何时使用 AI 的政策，包括组织允许用户与生成式 AI 共享哪些信息、访问控制指南、合规要求以及如何报告违规行为
- 开展影响评估，以收集信息、识别和量化使用 AI 的裨益和风险

2 增加对安全和隐私的可见性和控制

- 记录所有连接（包括到 AI 应用的连接），以持续监控用户活动、AI 工具使用情况和数据访问模式，以检测异常
- 发现存在哪些影子 IT（包括 AI 工具）——并做出批准、阻止或叠加额外控制的决策
- 扫描 SaaS 应用配置以识别潜在的安全风险（例如，从已批准的应用向未授权的 AI 应用授予 OAuth 权限，带来数据泄露风险）

3 检查进出 AI 工具的数据，过滤可能危害知识产权、影响机密性或违反版权限制的任何内容

- 对用户与 AI 工具的互动方式应用安全控制（例如：阻止上传，防止复制/粘贴，并扫描和阻止敏感/专有数据的输入）
- 实施保障措施，阻止 AI 机器人抓取您的网站内容
- 仅在不可能使用其他控制措施的情况下，才彻底阻止 AI 工具。众所周知，用户总会找到变通方法，导致安全失控

4 控制对 AI 应用和基础设施的访问

- 确保对访问 AI 工具的每个用户和设备都进行严格的身份验证，以限定谁可以使用 AI 工具
- 实施基于身份的 Zero Trust 访问控制。应用最低特权来限制帐户被盗或内部威胁的潜在损害

5 优化成本并提高运营效率

- 通过分析和日志记录了解人们如何使用 AI 应用，以便您可以随着使用量的增加而对速率限制、缓存、请求重试和模型回退加以控制

保护您构建的内容

训练您的 AI 模型

AI 管道正在导致漏洞的范围扩大。但凭借在开发整个过程中保障安全的经验，我们对实现成功的因素有独到的见解。对于 AI 安全，最自然的起点就是您的模型。

作为 AI 应用的基础，用于训练 AI 模型的所有内容都将影响其输出。从一开始就考虑如何保护这些数据，以避免日后的负面影响。如果不加保护，就可能导致攻击面扩大，并在日后造成应用问题。

确保数据完整性的安全措施对于减少故意和意外的数据泄露至关重要。AI 管道中的安全风险可包括：

- **数据投毒：**恶意数据集影响结果并造成偏差
- **幻觉滥用：**威胁行为者使 AI 幻觉（即为了生成响应而虚构信息）合法化，以便恶意和非法的数据集影响输出结果

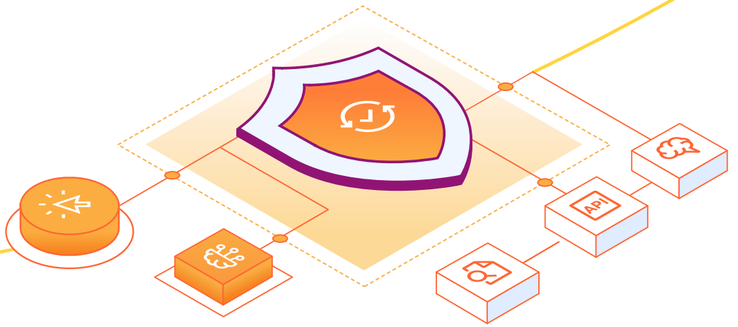
或者，如果您没有训练模型，那么您的内部 AI 将首先选择一个模型来执行任务。在这些情况下，您需要探索创作者如何创建和保护模型，因为它会影响到推理过程。

推理是 AI 模型训练之后的过程。模型训练得越好、调优程度越高，推理就越准确——尽管永远无法保证完美。即使经过充分训练的模型也会出现幻觉。

部署后的安全性

一旦构建并部署了内部 AI，您将需要保护其私有数据并确保访问安全。除了我们在本文中提出的建议（包括对每个用户提供令牌和实施速率限制）外，您还应该考虑：

- **管理配额：**使用限制来防止用户的 API 密钥泄露和共享
- **阻止某些自治系统编号 (ASN)：**防止攻击者向应用发送超负荷的流量
- **启用 Waiting Room 或质询用户：**使请求变得更加困难或更耗时，破坏攻击者的经济效益
- **构建和验证 API 模式：**通过识别和编目所有 API 端点来概述预期使用方式，然后列出所有特定参数和类型限制
- **分析查询的深度和复杂性：**帮助防止直接的 DoS 攻击和开发人员错误，维持源站正常运行，并按预期方式处理用户请求
- **建立以基于令牌的访问为中心的原则：**当令牌在中间件层或 API Gateway 验证时，防止访问遭到入侵



为您的生成式 AI 实验提供强大的威胁保护

从采用到实施，生成式 AI 实验的每个阶段都应该在风险最小或可容忍的情况下推进。借助从本文获得的知识，无论您的组织正在使用、构建还是规划在未来以某种形式采用 AI，您都有能力掌控自己的数字环境。

在采用新能力时感到犹豫是很自然的，但现有的资源可以给您信心，以安全地进行 AI 实验。在这些资源中，组织目前最需要的是 IT 和安全之间的连接纽带。它充当一条主线，通过与环境中的一切协作来简化复杂性，无处不在，并执行必要的安全、网络和开发功能。

通过这个连接纽带，您将对各种用例都充满信心，包括：

- 监管合规——能够检测和控制受监管数据的移动
- 对 SaaS 应用、影子 IT 和新兴 AI 工具中的敏感数据恢复可见性和控制
- 保护开发人员代码——检测和阻止上传和下载中的源代码。预防、发现和修复 SaaS 应用和云服务（包括代码存储库）中的错误配置

随着 AI 继续发展，不确定性将必然存在。因此，拥有 Cloudflare 这样的中流砥柱将大有裨益。

防范来自三种 LLM 的 AI 风险

根据使用情况，AI 给组织带来的风险暴露程度会有所不同。至关重要的是，了解与大型语言模型 (LLM) 使用和开发相关的各种风险，然后积极参与任何 LLM 部署。

LLM 的类型	主要风险
内部	对敏感数据和知识产权的访问
产品	声誉风险
公共	敏感数据泄露

规模、易用性和无缝集成

Cloudflare 的全球连通云将控制权交到您手中，并改善可见性和安全性——使 AI 实验安全且可扩展。更棒的是，我们的服务强化了一切，确保无需在用户体验和安全之间进行任何权衡。

考虑到大多数组织要么只使用 AI，要么同时使用并进行构建，利用 Cloudflare 意味着在 AI 项目上永不停滞。

- 我们的**全球网络**使您可以在任何需要的地方快速扩展和实施控制
- 我们**的易用性**使部署和管理用户AI 使用策略变得简单轻松
- 一个**可编程的架构**使您能够为正在构建的应用提供一个安全层，而不会影响用户使用 AI 的方式

Cloudflare 的全球连通云保护 AI 实验的每一个方面，特别是：

- 我们的 **Zero Trust 和安全访问服务边缘 (SASE)** 服务帮助减轻员工**使用**第三方 AI 工具时带来的风险
- 我们的**开发人员平台**帮助您的组织安全、高效地**构建**自己的 AI 工具和模型
- 在**利用 AI 的保护**方面，我们的平台利用 AI 和机器学习技术来构建威胁情报，然后用于在组织的 AI 实验中提供保护

	您如何使用 AI	您如何构建 AI
我们的全球网络规模	在任何地方一致地扩展和执行控制	加速推理、查询和缓存
我们管理简单性	单一控制平面，提供简单的部署和策略	丰富模板以便快速上手
我们统一、可编程的网络架构	叠加新的安全层而不影响您使用 AI 的方式	内置隐私保护和合规性

后续步骤

从保护组织使用 AI 的方式到保护您构建的 AI 应用，Cloudflare for AI 提供全面保障。通过我们的服务，您可以按任何顺序采用新功能，享受无限的互操作性和灵活的集成。

与专家讨论

如需了解更多信息，请访问 cloudflare.com/zh-cn

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

REV:BDES-6307.2024AUG19