技术白皮书

AI 渗透测试

大语言模型（LLM）安全评估

生成式 AI 正在迅速改变企业开发产品、自动化业务流程、支持员工以及与客户互动的方式。然而，随着大语言模型被集成到应用、数据管道、API、插件和自主智能体工作流中，它们也带来了全新而复杂的攻击面，传统安全测试可能无法全面覆盖。

当面临对抗性输入、不可信数据或遭篡改的上下文时，基于 LLM 的系统可能出现难以预测的行为。攻击者可能试图绕过安全控制、提取敏感信息、操纵输出、滥用关联工具，或诱导系统执行非预期操作。提示注入、越狱、智能体权限过大、不安全的输出处理、敏感信息泄露，以及检索增强生成（RAG）中的薄弱环节，都可能给业务、运营和数据保护带来显著风险。

Bitdefender 攻防安全服务提供定制化 LLM 安全评估，帮助企业在 AI 特有漏洞被利用之前，识别、验证并修复这些漏洞。每项评估均结合您的 AI 实现方式、运行逻辑和业务目标量身定制，帮助您清晰了解 LLM 应用的风险暴露，并持续提升其安全性。

评估范围

每个 AI 系统都各不相同。Bitdefender 会根据您的具体架构、使用场景、集成方式和风险状况，定制每项评估。

测试内容包括：

- 提示注入、越狱抵御能力及安全控制绕过
- 系统提示词泄露与敏感信息披露
- 智能体权限过大、非预期工具调用及关联系统滥用
- 检索增强生成（RAG）、向量存储及上下文操纵方面的薄弱环节
- 插件、API、连接器和数据流中的信任边界风险
- 日志记录、监控、检测、响应及 AI 业务逻辑方面的缺口

我们的评估覆盖整个 LLM 应用生态，包括应用逻辑、数据源、编排层、关联服务和用户交互入口。

Bitdefender

评估概览

Bitdefender LLM 安全评估在贴近真实攻击的对抗条件下，评估采用 AI 技术的系统。

通过自动化测试、人工对抗技术和受控漏洞利用，我们的顾问评估 LLM 应用如何处理输入、生成响应，以及如何与内部数据、外部工具、API 和业务流程交互，找出可能被攻击者用于操纵模型行为或滥用关联系统的安全缺口。

评估提供实用、以证据为依据的 AI 安全态势洞察，帮助安全、工程和风险团队理解漏洞影响，并确定修复优先级。

核心收益

- **全面掌握 AI 特有风险**
  覆盖提示词、交互接口、数据管道、API、插件、智能体及关联工作流。
- **先于攻击者发现漏洞**
  识别提示注入、越狱、数据泄露、智能体权限过大、不安全输出及工具滥用等风险。
- **验证 AI 部署安全性**
  通过自动化扫描和专家主导的人工对抗测试，为生产上线、扩容或更广泛的业务应用提供依据。
- **安全验证实际业务影响**
  在约定范围内，展示风险对数据、系统、用户和运营可能造成的实际影响。
- **支持更安全的设计决策**
  深入理解模型与应用、用户、工具和数据源的交互方式。
- **支持治理、审计与合规**
  提供严重程度评级、验证证据、复现步骤及修复建议。

测试方法

Bitdefender 将自动化测试的广泛覆盖与专家主导的人工对抗测试相结合，发现仅靠工具可能遗漏的薄弱环节。每项评估均根据您的 AI 架构、使用场景、集成方式和风险状况量身定制。

我们的测试分为四个主要阶段：

01 信息收集与范围界定

审查您的 AI 架构、数据流、模型接口、提示词、信任边界、关联 API、插件、智能体和检索系统，明确有针对性的测试范围。

02 对抗测试

模拟真实攻击，评估 LLM 应用如何应对操纵、不可信输入、提示注入、越狱尝试、数据提取、不安全输出及非预期工具调用。

03 受控漏洞利用

发现漏洞后，在约定范围内安全验证其潜在影响，判断这些薄弱环节是否可能导致数据暴露、未经授权的操作或业务逻辑滥用。

04 报告与修复

提供按风险评级的发现、支撑证据、复现步骤、业务影响分析及按优先级排序的修复建议，并在适用时对照公认的 AI 安全框架。

先于攻击者，守护您的 AI 系统

凭借攻防安全专长，以及对生成式 AI 应用构建、集成和受攻击方式的深入理解，Bitdefender 帮助企业识别与业务相关、可复现且可采取行动修复的漏洞。LLM 安全评估为安全、工程、治理和风险团队提供所需洞察，帮助验证控制措施、强化 AI 部署、降低滥用或数据暴露风险，让企业更有信心地采用生成式 AI。

联系我们

了解 LLM 安全评估服务，
欢迎联系我们。

申请安全评估

Bitdefender 中国

深圳总部

深圳市龙华区环智中心 D 座 1111

电话：4000-132-568

邮箱：sales@bitdefender-cn.com

微信扫一扫，关注我们

Bitdefender 是网络安全领域的领导者，为全球提供一流的威胁预防、检测与响应方案，守护数百万个人、企业和政府环境。作为业内备受信赖的安全专家，Bitdefender 致力于消除威胁、保护隐私与数据，并增强网络韧性。公司持续投入研发：Bitdefender Labs 每分钟发现 400 多种新威胁，每天验证约 400 亿次威胁查询，并在反恶意软件、物联网安全、行为分析和人工智能领域不断创新。其技术获全球 150 多家知名技术品牌授权采用。Bitdefender 成立于 2001 年，客户遍布 170 多个国家，并在全球设有办事处。

© 2026 Bitdefender。保留所有权利。文中商标、商号及产品归各自所有者所有。