安全+
SECURITY
2025/07 总第 065
技术版 ▶▶ 与安全人士分享技术心得 Share technique experience with security professionals

绿盟科技连续18年参展RSA大会，AI安全成果集中亮相

绿盟科技推出《数据要素安全：新技术、新安全激活新质生产力》

干货 | 绿盟科技网络安全研究报告“全家桶”来了

大模型安全风险分析与防护架构

NSFOCUS

本期看点 HEADLINES

主办：绿盟科技
策划：《安全+》编委会
地址：北京市海淀区北洼路4号院绿盟科技园
邮编：100089
电话：(010)6843 8880-5462
传真：(010)6872 8708
网址：www.nsfocus.com

2025/07 总第 065

欢迎您来信nsmagazine@nsfocus.com 与我们交流，分享您的建议和评论。（《安全+》部分图片来源于网络）

安全+ SECURITY+

目录 CONTENTS

卷首语

AI浪潮席卷全球，驱动数字经济深度演进，也深刻重塑了网络安全的技术范式与产业格局。随着大模型、智能体等新型技术形态逐步走向落地应用，安全问题不再局限于“传统边界”，而是渗透进从算法、模型到数据和交互的每一个环节。面对AI驱动的“新质生产力”，我们必须以系统性安全思维，为未来构建更加智能、弹性与可信的安全能力体系。

在2025年RSA大会上，绿盟科技连续第18年登上全球网络安全舞台，集中展示AI安全领域的核心成果，系统呈现我们在“AI+安全”融合方向上的深度探索。从AI大模型安全威胁矩阵V2.0到AI红队测试，从AI风险评估工具AI-Scan到智能安全运营平台ISOP，我们致力于构建贯穿大模型全生命周期的立体防护体系，以AI技术驱动网络安全的范式变革。更重要的是，我们正在推动从“AI赋能安全”迈向“安全治理AI”的双向路径，让每一次技术突破都服务于更强韧的数字生态。

与此同时，数据要素作为AI时代的“新石油”，其安全性关乎整个经济体系的高效运行与国家治理能力的现代化。在绿盟科技重磅推出的《数据要素安全：新技术、新安全激活新质生产力》中，我们系统提出了数据可信确权、可控流通、智能治理的全景方案，助力企业构建面向未来的数据安全底座。绿盟科技正在以平台化、智能化、标准化的战略路径，推动“数据安全即生产力”的理念落地生根。

安全不仅是技术问题，更是生态构建、规则演进和能力跃迁的问题。本期《安全+》将继续围绕AI安全、攻防实战、数据治理与政策趋势等关键议题展开探讨，凝聚行业智慧，激发技术灵感。我们期待与更多伙伴共建开放、智能、可信的数字安全新范式，为数字中国注入坚实的安全动能。

叶晓虎

绿盟科技连续18年参展RSA大会，AI安全成果集中亮相

绿盟科技 INTERNATIONAL BUSINESS 冀洁

摘要：在 2025 年 RSA 大会上，绿盟科技连续第 18 年亮相，集中展示 AI 安全领域的最新研究成果与技术方案，彰显其在“AI+安全”融合方向的深度布局。重点发布包括 AI 大模型安全威胁矩阵 V2.0、AI-Scan 风险评估工具、AI 红队测试、AI 安全赋能平台 NSFGPT、智能安全运营平台 ISOP 以及混合抗 DDoS 解决方案。绿盟科技在全球舞台上展现了中国网络安全企业的技术实力与持续创新能力，积极推动构建智能、高效、可信的数字安全体系。

关键词：AI 大模型安全 红队测试 RSA 大会 智能安全运营 抗 DDoS 解决方案

4 月 28 日至 5 月 1 日，全球网络安全领域的年度盛会——RSA Conference（RSAC）在美国旧金山盛大启幕。本届大会以“Many Voices. One Community”（多元声音，共同社区）为主题，旨在通过多元化的交流与合作，共同应对日益复杂的网络安全挑战。

自 2008 年起，绿盟科技已连续 18 年参与 RSA 大会。在全球网络安全持续演进的语境下，这份长期出现在国际舞台的坚持，既是一次次面向世界的技术答卷，也是一家中国网络安全企业对全球趋势的主动参与。

今年，绿盟科技聚焦 AI 时代的安全挑战，集中展示最新研究成果与前沿方案，与全球同行共同探讨智能安全的落地路径。

AI+ 安全融合重塑格局，绿盟科技创新成果集中展示

面对持续演进的威胁形态，绿盟科技围绕“AI+ 安全”战略，系统构建智能驱动的产品体系与能力矩阵。本次 RSA 大会，绿盟科技集中展示多项 AI 安全领域的关键成果，涵盖 AI 安全平台、大模型风险评估、AI 红队测试、智能运营、抗 D 防护等核心模块，体现出在安全技术与能力演进上的持续突破。

绿盟 AI 大模型安全威胁矩阵

随着大模型时代下各类业务应用形态的普及，传统的安全防御检测体系已无法满足 AI 系统应用防护的新需求。AISS 大模型安全社区本次升级了 AI 大模型安全威胁矩阵 V2.0 版本，对大模型安全进行了更加全面且细致地威胁建模。从基座安全、数据安全、模型安全、应用安全、身份安全以及大模型全生命周期出发，以多层次、立体化、全方位的视角来探索如何构建大模型安全防护体系。

绿盟 AI 大模型风险评估工具（AI-Scan）

为企业 AI 大模型的应用实践提供一个全面、深入的安全风险评估防线。该工具不仅涵盖了多种商业和开源大模型，还拥有迅速适配新兴大模型的能力，确保企业在采纳最新 AI 技术时的安全性。基于专家团队精心筛选和校准的测试用例库，它能够迅速且高效

地识别出内容安全和对抗安全的潜在威胁，同时，配备了专业的风险处理建议，为企业构建一道坚实的安全防护屏障。绿盟科技的这一创新工具将协助企业更好地理解和管理 AI 大模型带来的复杂安全挑战，确保企业在 AI 浪潮中乘风破浪，稳健前行。

绿盟 AI 红队测试（AI Red Teaming）

针对 AI 大模型应用潜在的安全风险提供全面评估，包括模型本身的安全性、在线系统的业务与服务安全，以及智能体运行环境的安全测试。其核心测试涵盖模型越狱、提示词注入等对抗性风险，帮助识别恶意操纵和数据泄露的可能性；同时，从应用角度进行系统安全和业务安全的测试，确保服务在实际运行环境中的安全稳健性；并进一步评估智能体的接口调用安全、运行环境安全、数据安全与网络安全。通过全方位的测试服务，绿盟 AI 红队测试为高风险场景、强监管行业和需要跨境合规审查的 AI 大模型应用提供了强有力的安全保障。

绿盟风云卫 AI 安全能力平台（NSFGPT）

集绿盟科技多年人工智能与机器学习研究经验、攻防知识与威胁情报积累、实战化专家能力于一体的 AI 安全赋能平台。平台内置多种大小模型、知识库、情报库，支持本地安全知识应用，并以智能体中心的形式场景化赋能各类安全产品及服务，实现模型能力拓展。通过将 AI 能力赋能安全产品与服务，平台可应用覆盖安全运营、检测响应、攻防对抗、知识问答等各类典型客户场景，实现网络安全智能化。

绿盟智能安全运营平台（ISOP）

在原有 XDR 能力基础上，与绿盟风云卫 AI 安全能力平台深度融合，实现基于 AI 的告警降噪，自主调查，未知威胁检测，行为基线建立等能力；通过 AI 辅助，实现检测、分析、研判、调查、响应等全流程 7*24 小时自主值守能力。通过创新 LUI 能力，支持语音及文字交互式生成态势感知界面。并发布多个安全智能体，覆盖检测、运营等方向，场景化地为客户提供安全运营能力。

绿盟混合抗 DDoS 解决方案

云地结合的智能抗 D 防护解决方案，可抵御各种类型的大流量 DDoS 攻击，并根据绿盟科技实验室持续不断地抗 D 研究成果和趋势观察更新产品技术和功能。最新版本的抗 D 更是加强了扫段攻击、DNS 随机子域名攻击和加密流量攻击检测与防护的能力，以前瞻性技术方案应对复杂化的 DDoS 攻击手段。

站在连续参展 RSA 第 18 年的节点上，绿盟科技不仅是技术展示者，更是全球安全生态的建设者。面向未来，我们将聚焦 AI 与安全深度融合，加快产品能力演进，推进标准化体系和合规能力建设，链接全球生态资源，共创可持续的数字安全未来。

RSAC 2025创新沙盒冠军解读| ProjectDiscovery：开源社区与Nuclei结合的攻击面管理

绿盟科技 创新研究院 桑鸿庆

摘要：本文详细介绍了 2025 年 RSAC 创新沙盒冠军 ProjectDiscovery 公司，聚焦其在攻击面管理（ASM）领域的独特优势。ProjectDiscovery 成立于 2020 年，总部位于旧金山，团队规模 11~50 人，依托强大的开源社区和云平台，提供高效的资产发现与漏洞扫描工具。文章重点解读了其核心产品与技术，特别是基于开源社区打造的 Nuclei、Httpx 和 Subfinder 工具，以及 AI 辅助的模板和资产标签自动生成能力。ProjectDiscovery 通过持续的资产监控与社区协作，推动攻击面管理工具的敏捷创新，满足现代企业对动态安全的需求。文章同时分析了该公司模式的优势与不足，展望其商业模式的稳定性和市场竞争力。

关键词：攻击面管理（ASM）开源社区 Nuclei 漏洞扫描器 AI 自动化模板生成 持续资产监控

1. 公司简介

ProjectDiscovery 成立于 2020 年，是一家专注于攻击面管理（ASM）的网络安全公司，总部位于美国旧金山。专注于提供开源和基于云的安全工具，以简化安全工程师和开发者的工作流程。

1.1 团队情况

ProjectDiscovery 团队规模 11~50 人，核心成员主要来自于印度。创始人 Rishiraj Sharma 在公司成立初期通过 GitHub 上的开源项目与其他创始人相识。Sandeep Singh 担任 CTO，专注于技术创新，特别是在自动化漏洞检测和攻击面管理方面。凭借初创团队在安全研究和自动化技术方面的专长，带领公司取得了显著成就。

图 1.Rishiraj Sharma（CEO）和 Sandeep Singh（CTO）[1]

1.2 融资情况

ProjectDiscovery 已通过两轮融资筹集了约 2800 万美元。这些资金用于扩展其云平台和增强产品功能（数据来自 Crunchbase）。

融资轮次	时间	金额	领投机构
种子轮	2021-02-09	170 万美元	SignalFire
A 轮	2023-08-17	2500 万美元	CRV

2. 产品背景

攻击面管理（ASM）已从简单的资产枚举演变为一个复杂的过程，能够持续发现、分类和监控所有易受攻击的资产。现代组织面临不断扩展的数字足迹，涵盖传统互联网暴露系统、动态云环境和复杂的分布式服务。2019 年分析师拉波特首次提出攻击面管理（ASM，Attack Surface Management）的概念，2021 年 7 月，Gartner《Hype Cycle for Security Operations》报告将 ASM 列为新兴技术，引发国内热潮，细分为外部攻击面管理（EASM）和网络资产攻击面管理（CAASM）。EASM 从攻击者视角分析公网暴露资产；CAASM 则通过 API 集成，提供内外部资产可见性和风险管理，助力企业在复杂 IT 环境中有效应对威胁。

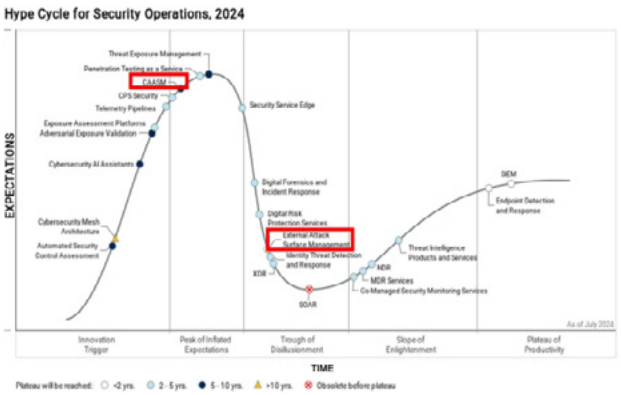


图 2. Gartner2024 年新兴技术成熟度曲线 [2]

传统漏洞管理工具存在明显局限，大多是较为固定的扫描模型，难以适应如今快速迭代的开发节奏。面对快速开发、动态基础设施和自动化攻击的新时代，安全团队亟需新的漏洞管理工具，真正提升防御效率与响应速度。ProjectDiscovery 通过结合成熟的开源技术和云原生功能，重新定义了资产攻击面管理。其平台通过对深度扫描和资产呈现，确保资产实时可见。简言之，它让安全团队能够以攻击者的视角看待组织的攻击面。

3. 方案特点

ProjectDiscovery 产品的 Slogan“大幅度减少资产扫描的时间、工具数量和开销”，其主要关注应用服务、内网资产、API、DNS、云、数据库资产发现，具体的实现架构和工具调用关系如下图所示：

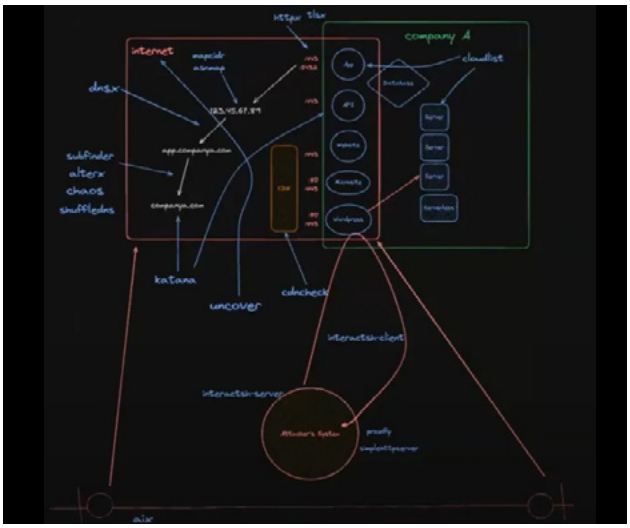


图 3. ProjectDiscovery 业务流程和工具调用关系示意图 [3]

3.1 持续的资产监控

ProjectDiscovery 是 SaaS 化的平台，持续监控组织中所有暴露于互联网的资产和服务，自动发现新主机、端点以及攻击面的变化。官网已开放试用，笔者输入了 ProjectDiscovery 官网域名做资产发现，平台的动化识别结果如图 4 所示。

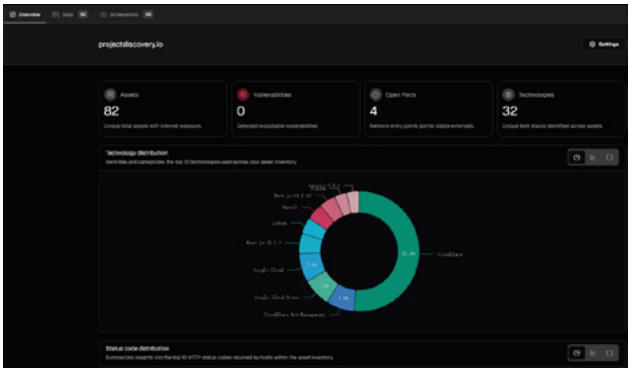


图 4. ProjectDiscovery 产品界面 [5]

从演示环境来看，资产识别涵盖了子域名、端口、状态码、IP 地址、ASN、CNAME 记录、使用技术栈、网页截图以及安全问题等多个维度。ProjectDiscovery 平台功能主要是对资产的数据的统计和搜索可视化，主要是针对互联网暴露面，数据的来源以主动探测发现为主，其对外宣传的内网资产风险发现功能并没有展示，笔者希望看到在路演中展示该功能。

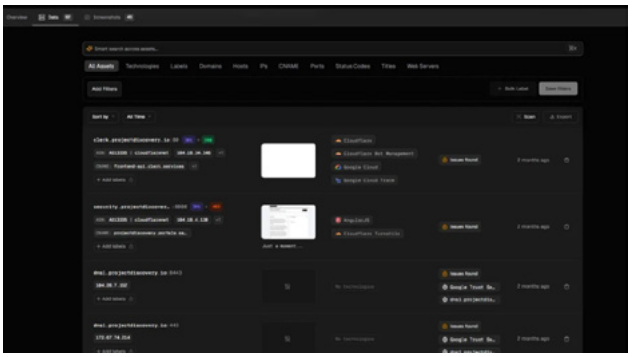


图 5.ProjectDiscoveryi 资产列表

3.2 开源社区集成

ProjectDiscovery 打造了一个由超过 10 万名工程师组成的蓬勃发展的全球社区，主要的工具包括：Nuclei, Httpx, Subfinder[6]。

3.2.1 Nuclei

Nuclei 是基于模板的可定制的漏洞扫描器，依托全球安全社区的支持，基于简洁的 YAML DSL 构建，来识别资产和脆弱性。它能够检测应用程序、API、网络、DNS 及云配置中的漏洞，目前项目在 GitHub 上已收获 22.9K Star，共计 9000 多个 Nuclei 模板。Nuclei 确保扫描速度快、结果精准，并与现实世界攻击者的行为保持一致。

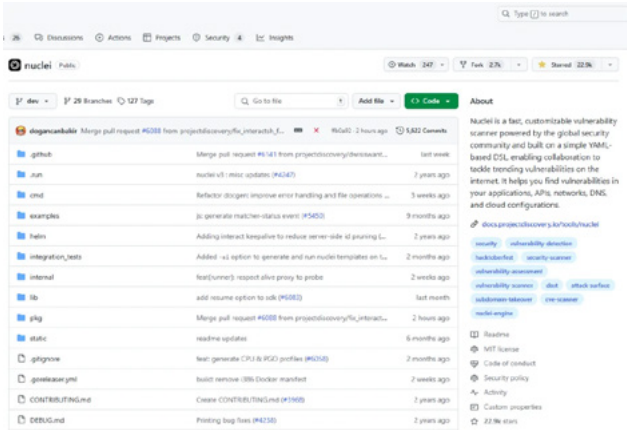


图 6.Nuclei 资产脆弱性探测工具

Nuclei 借助开源社区和自定义模板构建了强大的漏洞扫描生态，这是 ProjectDiscovery 的一大亮点，但笔者在实际并发测试中发现，尽管其相较传统的暴力扫描方式效率更高，在面对大规模资产探测时整体表现仍略显不足，推测其对丰富模板的支持在一定程度会影响了扫描效率，还是需要一些调度策略和扫描模式的支持。

3.2.2 Httpx

Httpx 是一款专注于 Web 资产探测的高性能工具，当前 Github

Star 数 8.4K。主要针对 HTTP 协议的服务探测，支持多种指纹识别与响应信息提取功能。能够高效获取状态码、标题、IP、TLS 信息、Favicon 哈希、截图、证书指纹等关键数据，尤其适合用于资产测绘和指纹识别前置环节，是一个强大的 Web 服务探测工具。

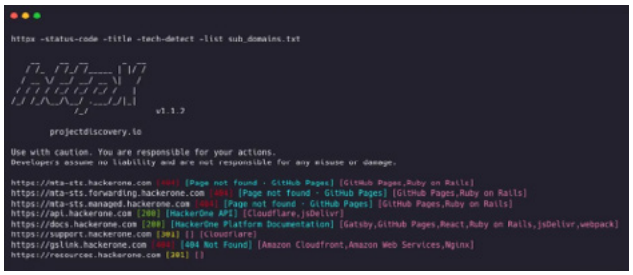


图 7.Httpx HTTP 探测工具

3.2.3 Subfinder

Subfinder 是一款专注于子域名发现的工具，旨在通过被动数据源快速、隐秘地枚举目标网站的有效子域名。其设计理念是简单高效，架构模块化，专注于被动子域名枚举这一核心功能。通过从在线被动数据源（如 DNS 记录、证书、搜索引擎、第三方 API 等）收集信息，发现目标网站的子域名。Subfinder 设计思路是避免主动扫描目标网络，确保操作的隐秘性，降低被检测的风险。

图 8.Subfinder 子域名发现工具

3.3 AI 辅助的内容生成

ProjectDiscovery 主要用 AI 辅助完成 Nuclei 检测模板和资产标签的自动化生成。

3.3.1 自动化模板生成

ProjectDiscovery 功能通过 AI 模板编辑器显著提升了 Nuclei 模板生成效率与质量。首先利用无头浏览器和 ChatGPT 从 POC 的链接中提取技术细节（如脚本、路径、HTTP 请求和 Payload），确保关键信息完整。通过 PDCP API 生成包含漏洞描述、元数据和匹配器的 Nuclei 模板，并由 TemplateMan API 优化元数据、CVSS 分类和格式统一。免费用户每日可进行 10 次 AI 请求，订阅和企业用户享有更高或无限制配额。笔者试用了该功能，生成最近的 Vite CVE-2025-30208 安全漏洞的 Nuclei 模板，效果还是很不错的，具体如图 9 所示。

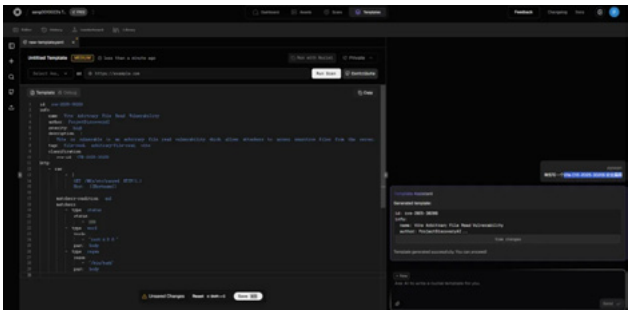


图 9.AI 生成 Nuclei 模板

3.3.2 自动化资产标签生成

ProjectDiscovery 支持 AI 驱动的资产标签功能，可自动分类并为资产添加上下文，帮助安全团队高效管理资产。官网描述该功能目前为早期测试版，首次标签可能较慢。通过分析资产元数据、DNS 记录、HTTP 响应及网页截图，系统智能分配描述性标签，如“登录页面”或“测试环境”，将原始数据转化为清晰的资产清单。

标签便于筛选与组织，清晰呈现攻击面，统一标准确保分类一致。新出现的资产可以自动获标签，保持清单实时准确。笔者在演示环境中增加标签是灰色锁定状态，应该还不支持试用，如图 10 所示。期待其路演介绍的实际展示。

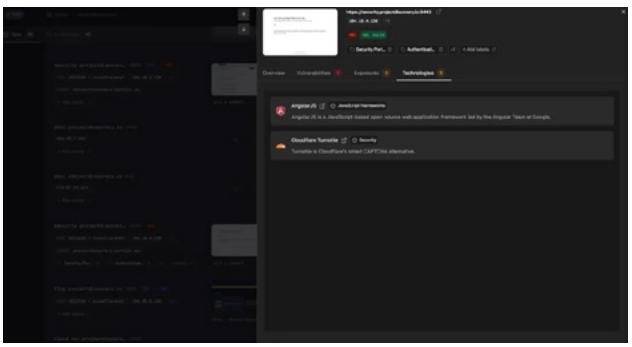


图 10.AI 生成资产标签

4. 总结

自 2019 年 Axonius[7] 凭借网络安全资产管理平台在 RSA 创新沙盒夺冠，其估值已飙升至 26 亿美元，并计划近期 IPO。2022 年，Sevco[8] 以数据融合为核心的资产管理平台入围，而 2025 年 ProjectDiscovery 凭借开源社区驱动的资产攻击面工具崭露头角，凸显资产攻击面管理（ASM）一直是 RSA 关注的热点。Axonius 和 Sevco 聚焦多源资产数据整合，提供全面可见性和安全管理；ProjectDiscovery 则通过开源社区打磨敏捷的漏洞发现能力。数据融合与社区驱动的两种路径，共同推动 ASM 领域的创新与发展。

ProjectDiscovery 特色在于融合开源驱动、全球社区协作与现代化平台，打造高效实用的漏洞管理工具，满足企业对动态、可视化安全体系的迫切需求，笔者认为 ProjectDiscovery 在技术上是有优势的，无疑是红蓝队都会使用的工具，但是模式上可能还缺少资产管理侧的功能，也就是从发现到响应的完整生命周期管理。同时，如何从开源生态活力转变成稳定商业模式可能也是需要考虑的，此外，应对其他厂商利用其开源成果构建竞品的挑战，将决定其在 ASM 市场的长期竞争力。

参考文献：

[1]https://www.crunchbase.com/organization/projectdiscovery-inc/financial_details.

[2]https://www.gartner.com/cn/newsroom/press-releases/2024-emerging-tech-hc.

[3]https://www.youtube.com/watch?v=cBkfk0VbvLw.

[4]https://projectdiscovery.io/#solutions.

[5]https://cloud.projectdiscovery.io/.

[6]https://github.com/projectdiscovery.

[7]https://cloud.tencent.com.cn/developer/article/1552167.

[8]https://cloud.tencent.com.cn/developer/article/2016448.

绿盟科技推出《数据要素安全：新技术、新安全激活新质生产力》

绿盟科技 创新研究院 陈佛忠

摘要：《数据要素安全：新技术、新安全激活新质生产力》由绿盟科技创新研究院多位数据安全专家联合撰写，系统阐述了数据要素安全的理论基础、技术路径及实践应用。书中围绕数据“可控流通、可信确权”以及生成式 AI 赋能的智能治理，深入分析数据要素安全的四大核心场景和多行业案例。该书适合企业决策者、技术开发者、数据流通推动者等，助力构建安全可信的数据流通体系，推动数字中国和数据要素化发展，助力国家治理能力现代化。

关键词：数据要素安全 可信确权 可控流通 生成式 AI 智能治理

在数字中国、数据要素化、人工智能驱动的新质生产力时代，数据的价值得到了前所未有的战略重视。从“数据上链、数据入表”，到“数据可控流通、可信确权”，再到以生成式 AI 赋能的智能治理，数据不再只是信息资源，更是重塑经济逻辑的关键生产要素。如何打破“数据孤岛”？如何实现数据的“可用不可见”？如何建立多方信任机制，让数据真正“流动起来”？这不仅是产业发展的关键命题，更是国家治理能力现代化的重要一环。

《数据要素安全：新技术、新安全激活新质生产力》正是在此背景下应运而生，由绿盟科技创新研究院多位数据安全专家深度参与撰写，系统呈现数据要素安全的理论基础、技术路径、实践应用和未来趋势，首次将“数据要素流通安全”作为独立主题进行全面剖析。

四大核心亮点

- 系统化安全框架

从宏观到微观，系统梳理数据要素安全的理论框架，深度解析数据要素演进与法规政策。

- 核心技术解析

聚焦数据安全自用、可信确权、可控流通、协同安全计算四大核心场景，提供实操性技术洞察。

- 多行业案例分析

结合多领域多行业的真实案例，为企业构建数据要素安全流通体系提供权威参考。

- 多层适读设计

内容涵盖趋势分析、技术细节与架构设计，满足初学者、技术专家和数据管理者的多样化需求。

谁适合读这本书?

- 企业决策者

系统性理解数据如何驱动业务创新与数智转型，为未来技术布局与安全治理提供方法参考。

- 数据流通推动者

掌握零信任、合成数据、数据脱敏等关键技术，降低跨域合作法律与技术风险。

- **技术开发者**

深入理解同态加密、差分隐私、可信执行环境等底层逻辑，结合开源工具实现落地实践。

- **行业研究者**

获得最新全球政策趋势、数据要素流通三部曲（确权—流通—定价）的理论与案例支持。

- **数据安全团队**

建立“可度量、可追责”的数据安全体系，推动组织安全能力从合规防护走向价值增值。

作者团队简介

刘文懋，绿盟科技首席创新官，中国计算机学会（CCF）理事、杰出会员。研究方向为云计算安全、数据要素安全、人工智能安全

等。出版《软件定义安全》《云原生安全》等著作，研究成果获北京市科学技术奖、中国计算机学会科技成果奖科技进步二等奖。

孟楠，长期从事信息通信网络安全、云计算安全等领域的科研和技术创新工作，ITU-T、ISO/IEC 信息安全注册专家，牵头制定了多项相关国际、国家和行业标准。

顾奇，绿盟科技安全研究员，广州大学方班企业导师。研究方向为数据安全、可观测性、密码学等，具备丰富的数据安全产品与方案设计经验。在 CCF-A/SCI 一区发表多篇高水平论文，曾获中国技术市场协会金桥奖一等奖、江苏省计算机学会科学技术奖三等奖。

陈佛忠，绿盟科技高级安全研究员，中国通信学会高级会员。研究方向为数据安全、云安全等，多次在知名会议上发表主题演讲。

高翔，绿盟科技安全研究员。研究方向为隐私计算、应用密码学。深度参与联邦学习开源项目，作为核心人员参与隐私计算产品开发。

王拓，绿盟科技安全研究员，研究方向为数据安全，专注于可信执行环境等隐私计算技术，探索大模型在数据安全领域的应用。作为核心人员参与了数据保险箱产品的设计与开发。

叶晓虎，绿盟科技集团首席技术官，先后担任国家火炬计划课题负责人、北京市下一代网络安全软件与系统工程技术研究中心主任、海淀区重大科技成果产业化负责人等。拥有 20 年信息安全管理的经验，多次参与国家级重大事件网络安全保障工作。

大咖推荐 品质背书

本书全面探讨了数据要素安全体系，涵盖法律法规、核心技术与实践案例。书中紧跟技术前沿，既深入介绍了传统数据安全技术，也对隐私计算、大模型等新兴技术进行了详细阐述，并提供了实践指导。无论是安全专业人士还是技术爱好者，都能从中获得实用知识与启发。

—— 任奎 浙江大学网络空间安全学院院长

从数据到数据要素，并非概念的简单升级，而是信息技术发展到新阶段的必然产物。数据要素安全也在不断探索新的边界与解法。本书对此进行了严谨、专业的探讨，值得大力推荐。

—— 白小勇 北京炼石网络技术有限公司创始人、CEO

数据的独特属性决定了其开放与共享的重要性。合规是释放数据价值的关键，安全则是保障数据价值流动的前提。本书从体系、技术、实践等维度系统阐述了数据要素安全的标准体系、合规现状、技术洞察及实践案例，是一部值得推荐的优秀作品。

—— 应志伟 海光信息副总裁

在生命科学领域，数据对推动科学发现至关重要，但其隐私性和敏感性常限制其价值的释放。绿盟科技数据安全团队在该领域开创性地探索了数据安全流转与利用的新模式，并获得多项荣誉。本书系统呈现了他们在数据要素安全领域的深厚理论与丰富实践，兼具深度与实用性，是关注数据驱动创新和数据安全的读者不可多得的前瞻性佳作。

—— 马俊才 中国科学院微生物研究所研究员
国家微生物科学数据中心主任

文悠博士的新书全面且深刻，深入探讨了数据要素安全的理论与实践，强调在技术快速发展的背景下如何确保数据的安全性和合规性，并提供了从数据生命周期管理到新技术应用的完整框架。对于数据安全专业人士、政策制定者以及关注数据隐私和安全的学者和学生，这本书是极具价值的重要参考。

—— 高志鹏 北京邮电大学计算机学院教授
中国计算机学会理事
数据治理发展委员会秘书长
中国通信学会数据安全专委会副主任

当前急需一本系统总结数据要素安全进展的图书，以供学术界和产业界参考。本书作者团队凭借多年经验，对数据要素安全领域的新技术进行了系统梳理和科学分类，并结合实际案例展示其应用。书中内容丰富扎实，前沿技术为科研人员提供灵感，先进技术为研发人员提供工具，生动案例为应用人员提供参考。我郑重推荐本书，期待它能促进数据要素的安全流通与价值增长。

—— 王宏志 哈尔滨工业大学长聘教授、博士生导师
海量数据计算研究中心主任
黑龙江省大数据科学与工程重点实验室主任

干货 | 绿盟科技网络安全研究报告“全家桶”来了

绿盟科技 能力中心 郭尧天

摘要：2025 年，网络安全攻防进入深水区，绿盟科技重磅发布多份网络安全研究报告，全面洞察低空经济、AI 攻防、APT 威胁、DDoS 演进等关键趋势。报告指出，无人机、AI 大模型、物联网设备已成为攻击者的新目标；APT 组织持续活跃，上海等核心城市遭遇集中打击。面对“云上战争”与“AI 博弈”，绿盟科技呼吁从通信加密、全生命周期防护到 AI 能力监管全面加强防御，同时以系列研究成果帮助企业识势、预警、布局，在数字化浪潮中守牢安全底线、抢占先机。

关键词：网络安全趋势 AI 攻防 无人机安全 僵尸网络 DDoS 演变

网络在变，攻击也在变，连黑客的套路都“卷”出了新高度。面对飞天的无人机、潜行的 APT、暗涌的 AI 安全风险，你真的看懂了吗? 你所看到的，仅仅是冰山一角，真正的安全真相，远比表面更复杂、更深远。

无人机飞了 2666 万小时，黑客也想“上天”

低空经济火热爆发，然而全球已记录 471 起无人机安全事件。从物流配送到应急救援，一旦通信链被劫持、数据被窃取、供应链被投毒——飞得越高，越要小心坠落。

防护建议：通信加密、全生命周期防护、零信任接入，一个都不能少。

AI 守门，AI 也在“挖洞”

2025 年告警降噪率高达 97%，安全运营效率提升 70%——AI 正成为安全防线的一部分。但别忘了，攻击者也在用 AI 生成钓鱼邮件、发起自动化攻击，甚至渗透大模型本身。

安全平衡点在于：对 AI 能力的监管、评估与实战测试。

新型僵尸网络 Hailbot 火力全开，直接轰瘫国产大模型 DeepSeek

中国遭到 DDoS 攻击量全球占比 34%，Hailbot + Rapperbot 联手，利用云资源发动 TB 级流量洪峰。云上安全形势已升级：不是“是否被打”，而是“打到你哪一块”。

全国暴露资产数：370 万台，最多的竟是摄像头和路由器

尤其是物联网资产，摄像头占比 50%+，路由器占比近 20%。这些都可能成为黑客入侵的“跳板”。

一句话提醒：资产暴露就是“在线裸奔”。

APT 组织总数破 620 个，上海成攻击“重灾区”

2024 年绿盟科技追踪到 296 起重大 APT 攻击事件，近半数攻击源来自境外。上海以 95% 占比高居国内 APT 攻击榜首——科研机构、制造企业、关键基础设施都在“暗影盯梢”下。

越是核心资产，越是黑客的目标。

2025 年，数字化与智能化浪潮加速奔涌，网络安全对抗进入深水区。绿盟科技锚定战略前沿，从低空经济到云端战场，从 AI 攻防到暗网态势，持续发布系列重磅研究成果，系统洞察安全格局与技术演进，助力用户识势、预警、布局，穿越数字迷雾，赢得先机。

报告合集

《网络安全 2025：冲刺 “十四五”》

报告系统梳理了 255 项国内外网络安全政策法规，全面分析了 APT 攻击、勒索软件、DDoS 演变趋势及 IPv6 安全等热点，帮助企业及政府理解并适应快速变化的网络安全治理新常态，提升战略决策能力。

《2025 网络安全趋势报告》

深入解读 2025 年网络安全行业十大趋势：安全大模型治理、AI 驱动红队攻防、可信数据空间构建、低空经济安全新体系等，并量化 AI 对安全运营效率的影响，助力企业前瞻布局、精细防控。

《低空经济启航，安全体系护航》

无人机产业迅速崛起的同时，安全问题层出不穷。绿盟科技联合工信部电子五所、西北工业大学深入调研发现通信劫持、数据泄露、供应链漏洞等风险，提出零信任架构和全生命周期防护策略，建成国内首个无人机攻防靶场，保障低空经济健康发展。

《高级威胁研究报告（2025 版）》

报告详细披露 2024 年全球 APT 攻击 296 起，涵盖战略欺骗、

零日漏洞利用、供应链污染等攻击手段，剖析 APT 攻击的战略性转向及精准打击模式，提升国家关键信息基础设施的防护能力。

《APT 组织研究年鉴》

基于知识图谱和大数据情报分析，报告系统梳理全球 620 个 APT 组织动态，首次详尽画像新增 55 个组织，分析 SSH、RDP 暴力破解占比达 91% 的入侵趋势，揭示上海等区域成为高密度 APT 攻击地区的原因。

《Botnet 趋势报告（2025 版）》

聚焦僵尸网络发展新趋势，披露中国遭受全球 34% 的 DDoS 攻击，Hailbot 等新型僵尸网络团伙兴起，成功瘫痪知名 AI 大模型服务，报告深入分析僵尸网络“带货模式”扩张现象，提供应对之道。

《DDoS 攻击威胁报告（2025 版）》

随着全球地缘政治冲突的不断升级以及网络空间冲突的兴起，DDoS 攻击的需求也在持续增长。攻击者不仅将其作为牟利工具，还将其视为一种战略手段，用于破坏关键基础设施或施加政治压力，DDoS 正逐步演变为政治诉求的重要载体。基于绿盟科技伏影实验室全球威胁狩猎系统监测与分析，正式发布《DDoS 攻击威胁报告（2025 版）》，系统性解构 DDoS 攻击趋势。

大模型安全规划：两类场景，五步走

绿盟科技 总工办 张睿

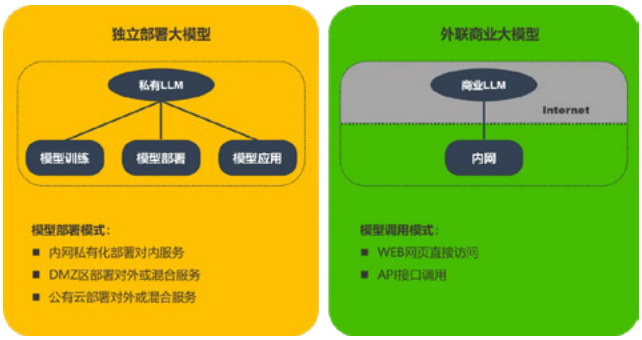
摘要:在“AI+”专项行动推动下,大模型已成为企业数字化转型的重要力量。本文提出“大模型安全规划”的两类场景与五阶段路径:在场景上，分别针对企业独立部署模型与外联商业模型制定差异化安全策略，以明确安全责任、控制数据外发风险；在路径上，基于内容安全护栏、Prompt 工程、模型微调、模型训练、数据安全五阶段构建纵深防御体系。规划强调将业务融合与风险可控并行推进，指导组织以实践导向开展模型安全全生命周期设计与建设，助力实现大模型在合规、安全前提下的落地应用。

关键词：大模型安全 场景规划 五阶段路径 风险防控

在国务院国资委“AI+”专项行动深化部署工作引导下，大模型被列为当前人工智能领域重点推进方向，要求加快构建步伐，推动关键核心技术取得实质进展。各类组织机构，尤其是国有企业，正积极谋划部署，加速推动大模型与实际业务的融合应用。作为一项新兴的信息技术应用形态，大模型因其通用性强、专用性广，正逐步扩展其在生产经营各环节中的实际影响。

1. 大模型两类场景

企业组织于顶层进行大模型安全规划时，需从两类应用场景分别设计，然后进行综合规划。两类场景分别为独立部署大模型场景，即企业组织机构自行部署大模型，能够控制模型的训练、部署、应用的部分或全部阶段；以及外联商业大模型场景，指企业组织机构通过互联网访问开放的商业模型，对模型不具有控制权。



先分类再整体的大模型安全规划思路，其优点主要从如下两个方面体现，首先，以模型所有权为基础，便于清晰划分安全责任。尤其针对第一类独立部署大模型的场景，组织机构基于大模型的生命周期，从训练、部署、应用分阶段梳理安全风险，分角色落实

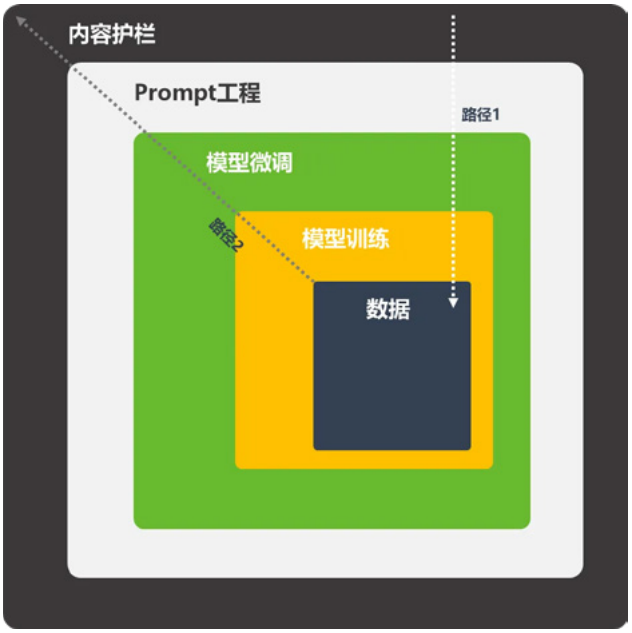
安全责任，从而依此进行安全设计。如以模型应用场景为例，对外发布大模型服务，除技术层面满足安全要求，还需重点考虑大模型合规备案、算法备案内容，所以在该阶段，组织机构安全部门于技术层面主导安全测试，尤其是上线发布前的综合测试，而在合规备案层面，往往需要业务部门主导，联合安全部门进行备案前安全自查、测试、报告等内容。

其次，单独剥离外联商业模型，利于大模型在风险受控的环境下应用，更符合风险管理原则与实践需要。因互联网商业大模型发展异常迅速，模型互联能力、应用场景丰富程度持续刷新，企业组织机构外联商业模型并提升自身业务能力、处理效率需求强烈，所以该场景一般需由安全部门进行总体设计，并区分 API 外联、Web 外联两类调用模式，尽快部署安全策略，防止组织内部商业敏感数据、个人信息乃至涉密数据的外发上传。尤其是该场景下的数据和文件外发风险高于传统网盘、论坛、社交软件，具有暴露面大、影响持久性长的特点。

在区分如上两类场景的基础上，综合安全设计时，需要协调数据安全于大模型应用场景的设计差异，既要考虑传统数据分类分级持续发挥作用，还需考虑模型外联数据源、自生成内容的安全和合规要求。此外，无论对于使用商业模型还是独立部署模型，也会因组织机构的风险偏好不同而呈现极大的特异性，其间对于平衡业务和安全的需求更加明显，且极大地提升业务部门以及数据所有者在整体大模型安全设计中的角色权重。

2. 五阶段安全规划

大模型安全规划除遵循安全技术措施同步规划、同步建设、同步使用，还可基于模型安全功能实践，分阶段实现内容安全护栏、Prompt 工程、模型微调、模型训练、数据安全。之所以分阶段实现，首先考虑了当前大模型应用场景异常活跃、监管要求持续细化，需要优先满足内容安全与合规要求，其次考虑模型应用自身及应用环境安全要求。内容安全包含了模型输入、输出两个方面，针对合规设计，需要重点考虑输出内容合规；此外，五阶段规划考虑了大模型技术以及相关安全技术的发展成熟度，同时兼顾工程成本和实现难度，所以从内容安全护栏开始，分阶段实现大模型安全工程化落地也更满足实践现状。伴随未来大模型及大模型安全技术不断发展成熟，五阶段安全规划可以平滑转换为大模型纵深防御架构，采取从中心到边缘的总体设计。



安全技术措施实现上，内容护栏主要为验证检查和过滤功能。在最终用户输入和大模型输出上执行所需安全检测、阻断控制措施。内容安全护栏充当用户和模型之间的代理和中介角色，使大模型能够专注于内容生成，而安全护栏则使应用程序安全、合规、可靠。

Prompt 工程是通过创建 Prompt 模板、检测并调整 Prompt，进而控制最终用户输入、输出的角色，以及控制数据类型、Token 长度等内容，Prompt 模板提供了预定义安全控制的能力。

模型微调是指基于既有安全需求，单独对模型执行额外微调，以保证模型在其指定应用领域、情景下按预期工作，尤其是满足国家、行业监管机构对于模型输出的要求。同时,也可基于业务场景、角色,进行微调。

模型训练从更早阶段开始，明确定义模型领域、角色、语气、场景，以进行针对性训练。如以网络安全领域为例，可以训练垂域模型，并且可进一步划分为攻击类场景和防守类场景，攻击场景大模型需以更严格的安全使用要求进行设计，防止模型滥用。以防守场景大模型为例，如设定“你是一个安全防守工程师，只回答防护建议，不回答攻击手段，基于你知道的回答，不确定的以及不知道的拒绝回答”。

若进一步追溯到大模型安全的上游环节，则必须对训练数据、微调数据进行安全设计，其涉及数据清洗、标注、脱敏、匿名化等工作，在分析模型的使用业务场景需求的前提下，定义数据范围、所有权、权限，涉及公开的需要预先评估数据风险，涉及商业推荐的，还需联合业务场景进行防歧视等算法合规层面的落实。

开源大模型应用的攻击面分析：云上LLM数据泄露风险研究系列（三）

绿盟科技 创新研究院 浦明

摘要：本研究聚焦开源大模型应用在“上云”过程中的数据泄露风险，指出便捷部署方式，虽加速技术落地，却带来 API 暴露、权限失效、密钥泄露等攻击面问题。研究选取多个热门开源应用（如 AnythingLLM、RAGFlow、Langflow、Dify 等），结合区域与云平台分布，揭示其全球部署现状及潜在威胁，呼吁开发者在快速迭代中加强安全防护，防范大模型生态中的数据泄露隐患。

关键词：数据泄露 大模型 大模型应用 LLM 攻击面分析 云上风险发现

往期回顾：

LLM 数据泄露风险专题研究文章：

《云上 LLM 数据泄露风险研究系列（一）：基于向量数据库的攻击面分析》

《云上 LLM 数据泄露风险研究系列（二）：基于 LLMOps 平台的攻击面分析》

1. 概述

本系列前两篇文章深入探讨了向量数据库和 LLMOps 在全球的暴露面及攻击面，本文作为第三篇，将重点关注当前主流大模型应用的安全风险。如今，大模型上云趋势明显，大多数大模型应用都可通过 Docker 快速一键部署，这种“一键上云”的便利虽然加速了技术落地，但也同时埋下了不少安全隐患，如未授权 API 接口调用、形同虚设的访问控制与权限失效、N Day 漏洞的再次利用问题等，均可导致用户隐私数据、机密信息大规模泄露。具体而言，攻击者可通过盗取大模型应用系统凭证、模型密钥、拦截聊天记录、污染训练数据等多重手段发起攻击。

本文我们依然从攻击面角度出发，对大模型应用中可能存在攻击的环节以及造成的实际危害进行分析，并给出缓解措施，希望可以通过具体介绍让大模型使用人员重视大模型生态中的数据安全。引发进一步思考。

2. 开源大模型应用介绍

我们认为大模型应用按类型可粗略分为问答系统（如 ChatGPT）、编程开发助手（如 Copilot）、搜索引擎与信息检索（如 New Bing）、RAG 应用（如 Langchain、FastGPT）、LLM 应用开发框架（Langflow、Dify）、垂直领域应用（如 IBM Watson Health）这几类，鉴于本系列文章更聚焦于开源生态，再加上我们调研开源问答系统、RAG 应用、LLM 应用开发框架较多，因此这两者将作为本文研究的重点。

我们调研发现，当前业界较为受欢迎的开源大模型应用如表 1 所示：

表 1 - 开源大模型应用基本信息

LLM应用	应用类型	Github链接	Star数	API
LangChain	RAG应用	https://github.com/langchain-ai/langchain	103k	不支持
AnythingLLM	RAG应用	https://github.com/Mintplex-Labs/anything-llm	40.9k	支持
FastGPT	RAG应用	https://github.com/labring/FastGPT	23.7k	支持
RAGFlow	RAG应用	https://github.com/infiniflow/ragflow	44.4k	支持
Vanna	RAG应用	https://github.com/vanna-ai/vanna	14k	支持
Langflow	LLM应用开发框架	https://github.com/langflow-ai/langflow	51.3k	支持
Dify	LLM应用开发框架	https://github.com/langgenius/dify	81.7k	支持
open-webui	LLM应用开发框架	https://github.com/open-webui/open-webui	82.9k	支持
NextChat	问答系统	https://github.com/ChatGPT-NextWeb/NextChat	81.9k	支持
Auto-GPT	问答系统	https://github.com/Significant-Gravitas/AutoGPT	173k	支持
ChatGPT Web	问答系统	https://github.com/Chanzhaoyu/chatgpt-web	31.9k	支持

3．开源大模型应用暴露面分析

我们针对上述小节中常见的大模型应用进行了测绘分析，重点为地区分布情况和所属云厂商两个维度，如下所示：

AnythingLLM

AnythingLLM 是由 Mintplex Labs 开源的一个全栈大模型聊天应用，该应用使用了现有商业大模型或开源大模型，再结合向量数据库解决方案以构建一个私有 ChatGPT，该应用可以本地运行，也可以远程托管，并能够与用户进行智能聊天。

Github 链接 :https://github.com/Mintplex-Labs/anything-llm

图 1 展示了 AnythingLLM 应用的全球区域分布与云厂商分布情况。AnythingLLM 应用全球服务数量超 6600 个，区域分布高度集中。前五大市场依次为中国、美国、德国、中国台湾省和新加坡，合计占比 84%。云服务部署呈现多元化特征：AWS 以 60% 占比

居首，阿里云（16%）与腾讯云（9%）次之，其余 15% 由其他云服务商构成。数据表明，尽管头部云厂商占据主要份额，但仍有相当比例的开发者选择非主流云平台进行部署。

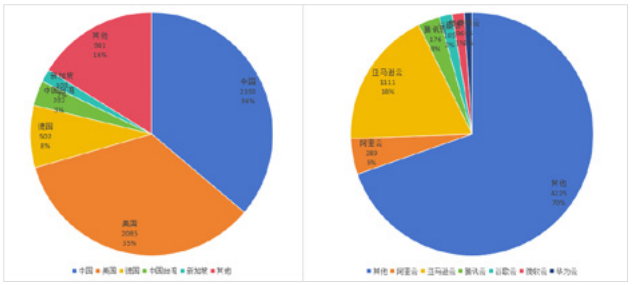


图 1 AnythingLLM 全球区域分布与云厂商分布情况

RAGflow

RAGFlow [9] 是一款基于深度文档理解构建的开源 RAG（Retrieval-Augmented Generation）应用。RAGFlow 可以为各种规模的企业及个人提供一套精简的 RAG 工作流程，结合大语言模型（LLM）针对用户各类复杂格式数据提供可靠的问答以及有理有据的引用。

Github 链接：https://github.com/infiniflow/ragflow

图 2 展示了 RAGFlow 应用的全球区域分布与云厂商分布情况。其中，RAGFlow 应用全球服务数量达 4800+ 个，区域分布呈现显著集中态势。前五大市场以中国（3200+ 次）、美国、德国、中国香港和南非为主，五地合计占比达 87%。云服务部署格局中，阿里云以 57% 的绝对优势领跑，腾讯云（18%）与 AWS（9%）分列二三位，其余 16% 由其他云平台承接。数据反映出中国市场在该应用中的核心地位，同时云服务生态虽以头部厂商为主，但长尾市场仍存在多元化部署空间。

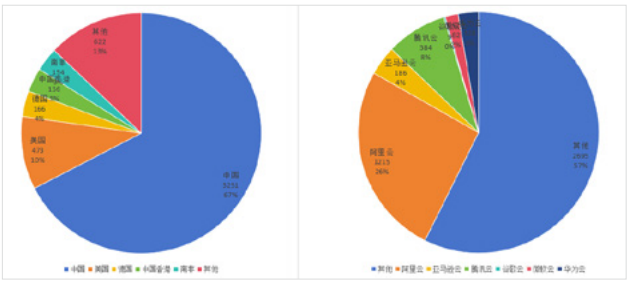


图 2 RAGflow 全球区域分布与云厂商分布情况

Langflow

Langflow 是一个强大的 AI 代理和工作流构建工具，提供可视化拖拽界面简化开发流程，内置 API 服务器可将代理快速部署为可调用端点，并开箱即用地支持主流大模型、向量数据库及丰富 AI 工具生态，实现无缝集成与快速部署。

Github 链接：https://github.com/langflow-ai/langflow

图 3 展示了 Langflow 应用的全球区域分布与云厂商分布情况。Langflow 应用全球服务数量约 2500 个，区域分布呈现明显区域性特征。前五大市场依次为美国、德国、印尼、英国和巴西，五国合计占比 63%，中国用户占比显著偏低，反映其在国内市场渗透度较弱。云服务部署集中度较高：AWS 以 58% 的绝对优势主导，谷歌云（19%）与微软云（16%）构成第二梯队，剩余 7% 由其他

云平台覆盖。数据凸显该应用在欧美及新兴市场的活跃度，同时反映出云服务生态仍由国际头部厂商把控的竞争格局。

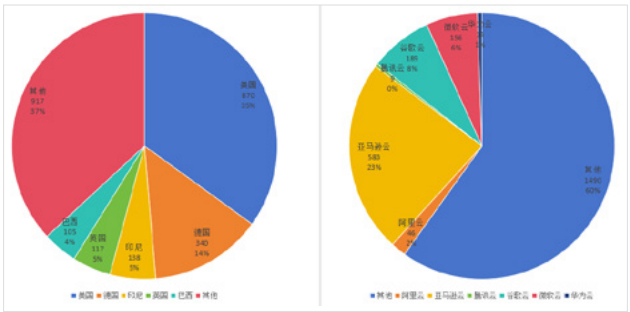


图 3 Langflow 全球区域分布与云厂商分布情况

Dify

Dify 是一个开源的 LLM 应用开发平台。其有较直观的界面并集成了智能 AI 工作流、RAG 管道、智能体能力、模型管理、可观测性等功能，有助于用户快速从原型开发过渡到生产部署。

Github 链接：https://github.com/langgenius/dify

图 4 展示了 Dify 应用的全球区域分布与云厂商分布情况。Dify 应用全球服务数量达 58000+ 个，呈现显著规模优势。区域分布高度集中：前五大市场为中国（30439 次，占比 52%）、美国、日本、德国和新加坡合计占比 86%，凸显中国市场的主导地位。云服务部署形成 " 一超多强 " 格局：阿里云以 43% 领跑市场，腾讯云（25%）与 AWS（17%）构成双巨头支撑，剩余 15% 由其他云平台覆盖。作为使用量远超同类组件的产品，Dify 展现出了中国市场强大的技术采纳能力。

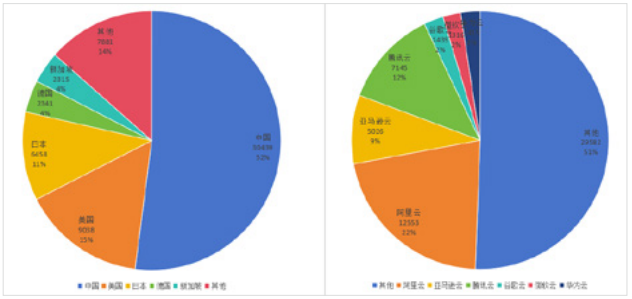


图 4 Dify 全球区域分布与云厂商分布情况

Open WebUI

Open WebUI 是一款可扩展、功能丰富且用户友好的自托管 AI 平台，支持完全离线运行。该平台兼容多种 LLM 运行环境（如 Ollama 及 OpenAI 标准 API），并内置 RAG 推理引擎，是强大的 AI 私有化部署解决方案。

Github 链接：https://github.com/open-webui/open-webui

图 5 展示了 Open WebUI 应用的全球区域分布与云厂商分布情况。Open WebUI 应用全球服务部署量达 104000+ 个，展现强劲市场渗透力。区域分布呈现双核驱动格局：中国（30349 个，占 29%）与美国（20547 个，占 20%）合计贡献近半数使用量，与德国、英国、中国香港共同构成前五大市场，五地合计占比 65%。云服务竞争呈三足鼎立态势:腾讯云（30%）以微弱优势领先阿里云（27%），AWS（19%）稳居第三梯队，剩余 24% 由多元云服务商分占。

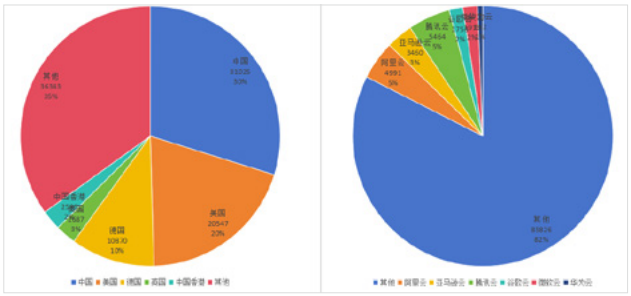


图 5 Open WebUI 全球区域分布与云厂商分布情况

NextChat

ChatGPT-Next-Web 是一款轻量化、可私有化部署的开源 ChatGPT 网页客户端，支持多模型 API 对接和本地数据存储，提供企业级对话 AI 快速集成方案。

Github 链接：https://github.com/ChatGPTNextWeb/NextChat

图 6 展示了 NextChat 应用的全球区域分布与云厂商分布情况。NextChat 应用全球服务部署量达 7600+ 个，区域分布中国（2348 个，占 29%）与美国（3454 个，占 20%）合计贡献 76%，与中国香港、新加坡、加拿大位居前五大市场。云服务部署呈现头部垄断态势：AWS 以 1,800+ 资产（45%）占据主导地位，阿里云（1,100+，28%）与腾讯云（800，20%）形成第二梯队，其余 7% 由其他云平台覆盖。

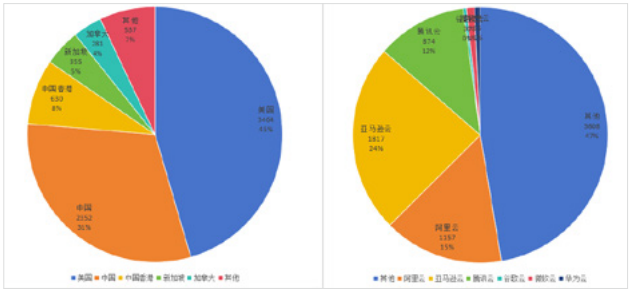


图 6 NextChat 全球区域分布与云厂商分布情况

4. 开源大模型应用数据泄露攻击面分析

我们对当前市场主流开源大模型应用架构研究后发现，当前市场中的开源大模型应用普遍存在 API 未授权访问风险，大模型应用在用户交互层与底层模型服务之间普遍采用 API 通信机制。部分应用存在接口鉴权机制缺失或配置缺陷，导致攻击者可绕过身份验证直接访问核心 API 接口。此类漏洞将直接暴露模型基础架构信息、应用凭证密钥、用户会话数据等敏感资产。攻击者利用该漏洞可实施模型资产窃取、会话数据爬取等高危操作，甚至通过密钥泄露实现横向渗透，下文我们将围绕以上风险对开源大模型应用的数据泄露攻击面进行分析。

4.1 模型基础信息泄露风险

经我们调研发现，部分聊天应用存在未授权 API 接口默认开放的安全隐患。这些接口可被任意访问并返回模型系统的基础信息，主要包括以下敏感数据：

(1) 向量数据库类型及配置参数；

(2) 模型版本信息；

(3) API 密钥调用状态等核心资产信息。

这些 API 接口已显露多重安全风险：首先，攻击者可利用获取的版本信息查询相关 CVE 漏洞数据库，精准定位已知漏洞进行渗透测试；其次，通过解析向量数据库类型及连接方式，攻击者可尝试构造未授权访问请求，实施数据窃取或注入攻击；更严重的是，API 密钥调用状态的泄露可能暴露系统脆弱环节，为攻击者提供横向渗透路径。

图 7 通过 API 未授权访问应用基本信息

图 8 通过暴露的基础信息对向量数据库进行未授权访问

4.2 LLM 应用凭证泄露风险

经我们分析发现，由于可未授权访问的 Web 聊天应用通常会将 API Token 等核心应用凭证内嵌在 Web 网页代码中，且较容易被发现，因此攻击者可借此发起以下两种类型攻击：

1. 模型资源滥用攻击

通过提取 Web 界面中的 API Token，攻击者可模拟合法身份调用官方 REST API 接口，对 AI 模型发起高频推理请求。此类“薅羊毛”攻击将造成算力资源劫持以及消耗模型服务配额，推高企业云计算成本的风险。

2. 权限体系穿透攻击

更严重的风险在于，部分应用会将管理员 API 密钥硬编码至前端代码。攻击者可逆向解析获取密钥后，直接访问后台管理接口，实施如隐蔽通道构建和权限架构破坏等行为。

图 9 通过 API 未授权访问获取大模型应用 Key

图 10 通过 API 未授权访问模型 swagger 文档

图 11 通过获取的大模型应用 Key 对应用本身进行操作

4.3 模型 Key 泄露风险

经研究发现，大模型应用开发框架（如 LangChain、FastGPT 等）在低代码编排场景中存在关键接口暴露风险。此类框架通过可视化流程（Flow）实现多模型链式调用，但若使用者未对 Flow 编排接口进行严格管控，攻击者可利用以下路径实施深度攻击：

攻击链：

阶段一：Flow 元数据探测，未鉴权 Flow API 访问 → 响应数

据分析 → 提取模型密钥 / 终端地址：

阶段二：横向权限穿透，将窃取的密钥注入自有应用，劫持原付费账户的模型调用权限：

阶段三：资源定向破坏，使用合法密钥发起高并发推理请求，触发模型服务商计费风控阈值或创建影子账户，结合泄露的管理密钥，在模型服务商侧注册傀儡账户实现持久化。

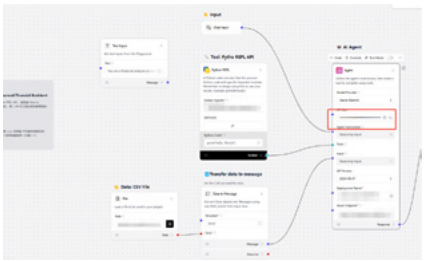


图 12 通过 API 未授权获取大模型 Key

4.4 模型聊天信息泄露风险

API 未授权访问问题也可能导致模型交互数据的系统性泄露，其攻击路径主要呈现为聊天记录直提取以及 LLM 凭证劫持

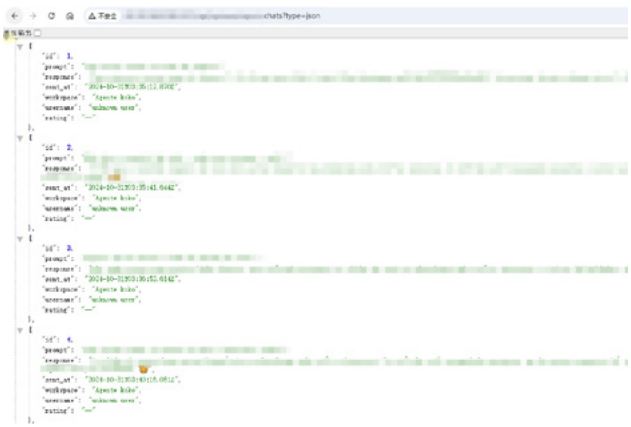


图 13 通过 API 未授权获取模型聊天记录

4.5 模型训练信息泄露

通过 API 未授权访问，攻击者可以访问模型训练数据，图 14 为大模型应用的训练数据页面，可以看出是通过自然语言生成 SQL 语句，虽然给使用者带来了便利，但若不对相应接口进行鉴权，训练数据则会被暴露于外，攻击者也可根据历史指令构造恶意语义化指令，从而绕过业务逻辑窃取敏感信息。

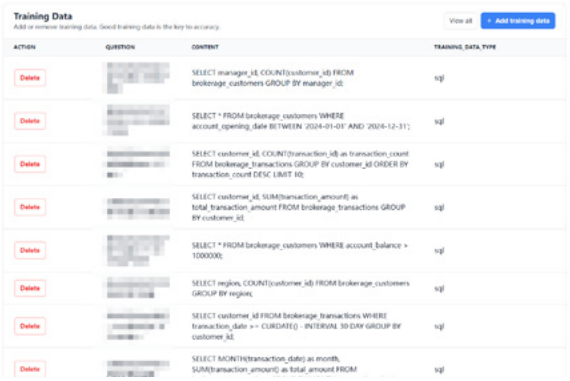


图 14 模型训练数据泄露

5. 建议防护措施

由于上述风险和攻击面均是由于不安全的 API 访问所引起的，因此我们需要从根源解决问题，建议应用部署者进行以下防护措施：

1. 强制启用 OAuth 2.0/SSO 等认证方案，对模型应用 Web 控制台实施 IP 白名单访问控制；

2. 针对管理类 API 接口实施双因素认证，按最小权限原则划分 API 访问等级；

3. 针对模型开发框架类应用建议对 Flow 接口实施严格鉴权，最大幅度减少其 API 暴露面；

4. 针对语义绕过进行严格安全层过滤。

6. 绿盟科技创新研究院云上风险发现研究成果

绿盟科技创新研究院在云上风险发现和数据泄露领域已经开展了多年的研究。借助 Fusion 数据泄露侦察平台，我们已监测到数万个云端暴露资产存在未授权访问的情况，包括但不限于自建仓库、公有云对象存储、云盘、OLAP/OLTP 数据库、大模型组件，以及各类存储中间件等，具体研究内容可参考包括但不限于 DevSecOps 组件、自建仓库、公有云对象存储、云盘、OLAP/OLTP 数据库、大模型组件以及各类存储中间件等，具体研究内容可参考《2023 公有云安全风险分析报告》[1]，《2024 上半年全球云上数据泄露风险分析报告》[2]，《全球云上数据泄露风险分析简报》第一期至第五期[3-7]。

Fusion 是由绿盟科技创新研究院研发的一款面向数据泄露测绘的创新产品，集探测、识别、泄露数据侦察于一体，针对互联网中暴露的泛云组件进行测绘，识别组件关联的组织机构和组件风险的影响面，实现自动化的资产探测、风险发现、泄露数据分析、责任主体识别、数据泄露侦察全生命周期流程。

图 15 Fusion 能力全景图

Fusion 的云上风险事件发现组件具有如下主要特色能力：

资产扫描探测：通过多个分布式节点对目标网段 / 资产进行分布式扫描探测，同时获取外部平台相关资产进行融合，利用本地指纹知识库，实现目标区域云上资产探测与指纹标记；

资产风险发现：通过分布式任务管理机制对目标资产进行静态版本匹配和动态 PoC 验证的方式，实现快速获取目标资产的脆弱性暴露情况；

风险资产组织定位：利用网络资产信息定位其所属地区、行业以及责任主体，进而挖掘主体间存在的隐藏供应链关系及相关风险。

资产泄露数据分析：针对不同组件资产的泄露文件，结合大模型相关技术对泄露数据进行分析与挖掘，实现目标资产的敏感信息获取；

参考文献：

[1]《2023 公有云安全风险分析报告》https://book.yunzhan365.com/tkgd/qdvx/mobile/index.html 。

[2]《2024 上半年全球云上数据泄露风险分析报告》https://book.yunzhan365.com/tkgd/cltc/mobile/index.html 。

[3] 全球云上数据泄露风险分析简报（第一期）https://book.yunzhan365.com/tkgd/sash/mobile/index.html。

[4] 全球云上数据泄露风险分析简报（第二期）https://book.yunzhan365.com/tkgd/bxgy/mobile/index.html。

[5] 全球云上数据泄露风险分析简报（第三期）https://book.yunzhan365.com/tkgd/xyih/mobile/index.html 。

[6] 全球云上数据泄露风险分析简报（第四期）https://book.yunzhan365.com/tkgd/xbin/mobile/index.html 。

[7] 全球云上数据泄露风险分析简报（第五期）https://book.yunzhan365.com/bookcase/wxjf/index.html。

开源大模型应用的攻击面分析：云上LLM数据泄露风险研究系列（四）

绿盟科技 创新研究院 浦明

摘要:本文系统分析了开源大模型推理软件的安全风险，重点揭示 Ollama、Fastchat、llama.cpp 等因 API 未授权访问、N-Day 漏洞（如 SSRF、目录遍历）导致模型数据泄露和远程攻击的高危场景。绿盟科技基于 Fusion 平台监测发现云上暴露面广泛，建议企业加强鉴权、隔离部署、漏洞修复与日志监控，共同防范大模型生态的数据安全威胁。

关键词：数据泄露 大模型 大模型推理软件 LLM 推理框架攻击面分析 云上风险发现

1. 概述

作为本系列的第四篇，本文聚焦大模型推理软件的安全风险。随着大模型上云趋势加速，尽管推理框架通常被视为底层基础设施（负责模型运行的资源调度与计算优化），但经我们的研究发现，部分开源推理框架，如 Fastchat、Ollama、llama.cpp 因配置缺陷或 N Day 漏洞暴露于公网，形成新型攻击面。例如：

- Fastchat Web 服务器被曝出有 SSRF 漏洞，攻击者可以访问原本无法访问的内部服务器资源数据，例如 AWS 的元数据凭证。
- Ollama 默认开放的 11434 端口导致攻击者可进行 API 未授权访问从而导致模型参数、训练数据被窃取；
- llama.cpp 的 RPC-server 历史漏洞（如 CVE-2024-42479）允许攻击者通过内存破坏实现远程代码执行，进而控制分布式集群节点；
- llama.cpp 的服务状态监控接口暴露可能泄露模型加载信息、实时推理请求等敏感数据。

本文从攻击链视角出发，系统分析大模型推理软件的关键风险环节、实际危害及缓解措施，旨在推动行业重视大模型生态中的数据安全与基础设施防护。

2. 开源大模型推理软件介绍

根据我们的调研，目前主流的开源大模型推理软件如表 1 所示，这些框架因其性能、易用性和社区支持度在开发者中广受欢迎：

表 1 - 开源大模型推理软件基本信息

LLM应用	Github链接	Star数	API
Ollama	https://github.com/ollama/ollama	139k	不支持
vllm	https://github.com/vllm-project/vllm	45.9k	支持
LightLLM	https://github.com/ModelTC/lightllm	3.2k	支持
OpenLLM	https://github.com/bentoml/OpenLLM	11.2k	支持
llama.cpp	https://github.com/ggml-org/llama.cpp		
Hugging-Face TGI	https://github.com/huggingface/text-generation-inference	10.1k	支持

GPT4ALL	https://github.com/nomic-ai/gpt4all	73.2k	支持
Fastchat	https://github.com/lm-sys/FastChat	38.6k	支持

3．开源大模型推理软件暴露面分析

我们针对上述小节中常见的大模型推理软件进行了测绘分析，重点为地区分布情况和所属云厂商两个维度，如下所示：

Ollama

Ollama 是一个轻量级开源框架，专注于简化大型语言模型（LLM）的本地化运行与管理，支持用户一键部署、交互和定制主流开源模型（如 llama、Mistral 等），无需复杂配置即可实现高性能本地推理。

Github 链接：https://github.com/ollama/ollama

图 1 展示了 Ollama 服务的全球区域分布与云厂商部署特征。Ollama 全球暴露面总量达 302748 个实例，区域分布呈现明显梯队化：美国以 92943 例（30.7%）居首，中国（21204 例，7.0%）、德国（19160 例，6.3%）分列二三位，前五大市场合计贡献全球部署量的 55.3%。云服务部署呈现高度集中化趋势，亚马逊云以 199383 例独占 65.8% 的绝对份额，与其他云厂商形成显著差距—其余厂商中占比最高的阿里云（3960 例）仅占 1.3%，而 " 其他 " 云平台以 94511 例（31.2%）成为第二大部署选择。数据表明，Ollama 的云生态呈现 " 一超多弱 " 格局，但开发者对非主流云平台仍保有相当程度的使用偏好，这与头部云厂商的技术锁定效应形成有趣对比。

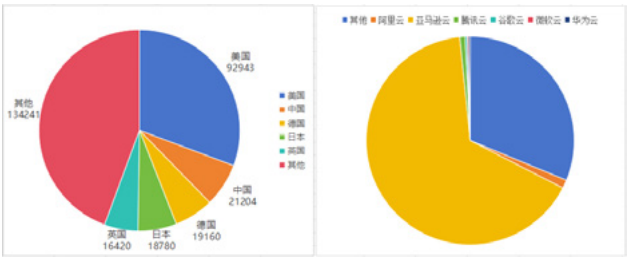


图 1 Ollama 全球区域分布与云厂商分布情况

Fastchat

FastChat 是一款专为大型语言模型（LLM）设计的开源框架，支持聊天机器人的全流程开发（训练、部署、评估），兼容主流开源模型与 OpenAI API，提供高效推理工具和交互界面，降低智能对话服务开发门槛。

Github 链接：https://github.com/lm-sys/FastChat

图 2 展示了 Fastchat 服务的全球部署情况及云平台分布特征。 Fastchat 全球暴露量共计 6450 个实例，区域分布呈现明显分化：中国（2403 例，37.3%）和美国（1941 例，30.1%）占据主导地位，合计贡献近 70% 的部署量，而德国、韩国、日本等市场占比均不足 5%，呈现“中美双核”格局。云服务部署方面，亚马逊云（1418 例，22.0%）位列第一，谷歌云（413 例，6.4%）和阿里云（675 例，10.5%）分列二三位，而腾讯云（326 例，5.1%）与微软云（145 例，2.2%）占比相对有限。值得注意的是，“其他”云平台（3,420 例，53.0%）占比过半，表明 Fastchat 开发者更倾向于选择非主流云服务商或私有化部署方案，与 Ollama 的集中化

趋势形成鲜明对比。数据表明，Fastchat 的云生态呈现多元化特征，头部云厂商尚未形成绝对垄断，开发者对中小云服务商或自建基础设施的依赖度较高。

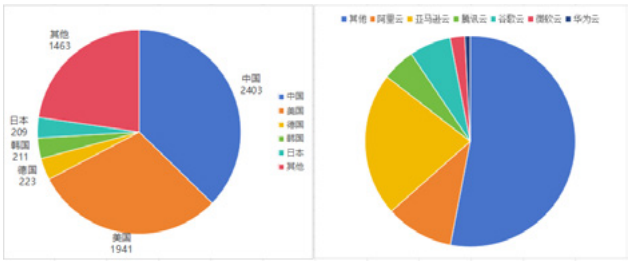


图 2 Fastchat 全球区域分布与云厂商分布情况

llama.cpp

llama.cpp 是一个高性能的 C/C++ 推理框架，专为在 CPU/ 边缘设备 上高效运行 Llama、Mistral 等开源大模型 而设计，支持量化（4-bit/5-bit/8-bit）和轻量化部署，实现低资源消耗下的快速推理。

Github 链接：https://github.com/ggml-org/llama.cpp

图 3 展示了 llama.cpp 服务的全球部署情况及云平台分布特征。 全球暴露量共计 3,880 个实例，区域分布呈现“美中德主导、长尾分散”的特点：美国（1118 例，28.8%）位居第一，中国（629 例，16.2%）和 德国（395 例，10.2%）紧随其后，前三大市场合计占比 55.2%，而日本、英国等地区占比均不足 5%，剩余 35.7% 的实例分布于全球其他区域。

在云服务部署方面，llama.cpp 展现出极强的去中心化趋势：“其他”类别（3183 例，82.0%）占据绝对主导，远超主流云厂商，表明开发者更倾向于使用 私有化部署、本地服务器或中小型云服务商。

主流云平台，亚马逊云（310 例，8.0%）占比最高，但份额远低于行业平均水平；谷歌云（128 例，3.3%）和 微软云（128 例，3.3%）并列第二，而 阿里云(67 例，1.7%) 和 腾讯云(59 例，1.5%) 占比极低，华为云（5 例，0.1%）几乎可忽略不计。

数据表明，llama.cpp 的部署模式与 Fastchat、Ollama 存在显著差异：

云依赖度低：82% 的实例运行在非主流云平台，说明其用户群体更偏好本地或自托管方案，而非公有云。

亚马逊云影响力减弱：仅占 8%，远低于 Ollama（65.8%）和 Fastchat（22%），反映 llama.cpp 的技术栈可能更适配轻量化、离线或边缘计算场景。

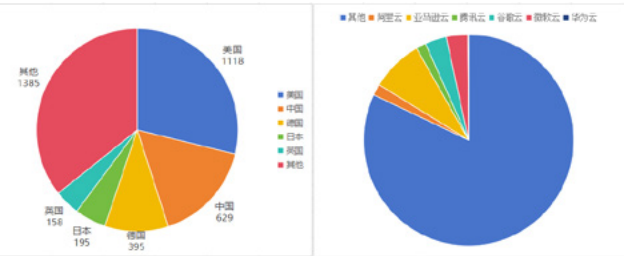


图 3 llama.cpp 全球区域分布与云厂商分布情况

四．开源大模型推理软件模型泄露攻击面分析

我们对当前市场主流开源大模型推理软件研究后发现模型推理框架普遍存在 API 未授权访问风险、大模型推理软件与模型应用基本一致，在用户交互层与框架服务间普遍采用 API 通信机制。并且存在接口鉴权机制缺失或配置缺陷，导致攻击者可绕过身份验证直接访问核心 API 接口。此类脆弱性配置将会导致模型推理框架中的模型基础架构信息、运行时信息、内部资源等敏感资产泄露。攻击者利用大模型推理软件的 N-Day 漏洞，如 SSRF、目录穿越

等漏洞可实施服务器内部资源窃取等高危操作，下文我们将围绕以上风险对开源大模型推理软件的数据泄露攻击面进行分析。

4.1 模型基础信息泄露风险

经我们调研发现，部分大模型推理软件存在未授权 API 接口默认开放的安全隐患。这些接口可被任意访问并返回模型系统的基础信息，包括所用模型配置详情及参数规模等敏感数据，这些 API 接口已显露多重安全风险，如攻击者可利用获取的版本信息查询相关 CVE 漏洞数据库，精准定位已知漏洞进行渗透测试。

图 4 通过 API 未授权访问 Ollama 使用模型基本信息

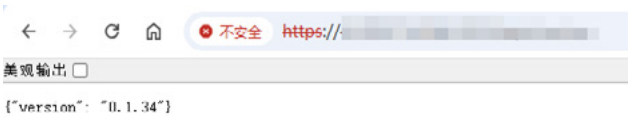


图 5 通过 API 未授权访问 Ollama 的版本信息

4.2 模型运行时信息泄露风险

我们的安全研究表明，开源大模型推理软件 llama.cpp 在其服务器模式的 Restful API 实现中存在安全缺陷。该框架采用 Slots（槽位）机制作为并行处理能力的核心支撑，每个槽位对应独立的推理上下文，可同时服务多个用户请求并实现细粒度监控。然而，其 API 接口未配置身份认证及权限控制模块，攻击者可借此进行未授权访问，进而引发敏感数据泄露攻击，如通过非法调用 /slots 监控接口，攻击者可完整获取槽位运行状态、推理任务队列、用户提问记录等敏感数据。结合历史请求中的企业知识库问答记录，可实施训练数据窃取。

图 6 通过 API 未授权访问 llama.cpp 的 slots 信息

4.3 模型服务器内部资源和数据泄露风险

我们调研到，开源大模型推理软件如 OpenLLM 与 FastChat 都不同程度存在目录遍历、SSRF 等 N-Day 漏洞，进而导致服务器可能存在以下风险点：

本地敏感文件暴露（OpenLLM CVE-2024-8982）：通过 LFI 漏洞的目录遍历，攻击者可读取服务器本地的配置文件（如 /etc/passwd、/proc/self/environ）、密钥文件（如 SSH 私钥）、日志文件等；此外，若模型文件存储路径暴露，攻击者可通过 LFE 漏洞窃取专有模型参数。

内部网络探测与横向渗透：利用 FastChat CVE-2024-12376 SSRF 漏洞，攻击者可构造恶意请求访问内部服务（如数据库管理接口、Kubernetes API Server、云服务元数据接口），从而获取云厂商临时凭证、容器集群配置或内网服务漏洞，为横向移动提供跳板。

五．建议防护措施

强制严格进行用户鉴权标准化认证：使用 OAuth 2.0/JWT，禁用匿名访问，默认拒绝未认证请求。

精细权限控制：基于 RBAC/ABAC 限制 API 访问范围。

输入验证与路径限制：对用户输入进行严格正则匹配，禁止 ../ 等目录遍历字符；使用白名单机制限制文件访问范围。

网络隔离与访问控制：将模型服务器部署于私有子网，禁用对外部元数据服务的访问；实施最小权限原则，限制服务账户权限。

日志监控与威胁检测：部署 AI 驱动的异常行为分析系统，实时监控 LFI/SSRF 攻击特征，并联动 WAF 拦截恶意请求。

漏洞响应与修复：参考 OWASP LLM Top 10 安全指南，定期进行红队演练，及时修复框架漏洞。

六．绿盟科技创新研究院云上风险发现研究成果

绿盟科技创新研究院在云上风险发现和数据泄漏领域已经开展了多年的研究。借助 Fusion 数据泄露侦察平台，我们已监测到数万个云端暴露资产存在未授权访问的情况，包括但不限于自建仓库、公有云对象存储、云盘、OLAP/OLTP 数据库、大模型组件，以及各类存储中间件等，具体研究内容可参考包括但不限于 DevSecOps 组件，自建仓库、公有云对象存储、云盘、OLAP/OLTP 数据库，大模型组件以及各类存储中间件等，具体研究内容可参考《2023 公有云安全风险分析报告》[1]，《2024 上半年全球云数据泄露风险分析报告》[2]，《全球云上数据泄露风险分析简报》第一期至第五期 [3-7]，云上 LLM 数据泄露风险研究系列 [8-10]。

Fusion 是由绿盟科技创新研究院研发的一款面向数据泄露测绘的创新产品，集探测、识别、泄露数据侦察于一体，针对互联网中暴露的泛云组件进行测绘，识别组件关联的组织机构和组件风险的影响面，实现自动化的资产探测、风险发现、泄露数据分析、责任主体识别、数据泄露侦察全生命周期流程。

图 7 Fusion 能力全景图

Fusion 的云上风险事件发现组件具有如下主要特色能力：

资产扫描探测：通过多个分布式节点对目标网段 / 资产进行分布式扫描探测，同时获取外部平台相关资产进行融合，利用本地指纹知识库，实现目标区域云上资产探测与指纹标记；

资产风险发现：通过分布式任务管理机制对目标资产进行静态版本匹配和动态 PoC 验证的方式，实现快速获取目标资产的脆弱性暴露情况；

风险资产组织定位：利用网络资产信息定位其所属地区、行业以及责任主体，进而挖掘主体间存在的隐藏供应链关系及相关风险。

资产泄露数据分析：针对不同组件资产的泄露文件，结合大模型相关技术对泄露数据进行分析与挖掘，实现目标资产的敏感信息获取；

当今数字化迅速发展的时代，数据安全问题越来越受到广泛关注。与此同时，随着云计算技术的普及和应用，企业也不可避免面临着云上数据泄露事件的频繁发生，为了提供公众和相关行业对数据安全的认知，我们计划定期发布有关云上数据泄露的分析报告，这些报告将以月报或双月报的形式呈现，内容涵盖最新的云上数据泄露案例分析、趋势洞察、数据保护最佳实践以及专家建议等。

如果读者对本文有任何意见或疑问，欢迎批评指正。如有合作意向请联系我们（邮箱 chenfozhong@nsfocus.com）。

参考文献：

[1]《2023 公有云安全风险分析报告》https://book.yunzhan365.com/tkgd/qdvx/mobile/index.html

[2]《2024 上半年全球云上数据泄露风险分析报告》https://book.yunzhan365.com/tkgd/cltc/mobile/index.html

[3] 全球云上数据泄露风险分析简报（第一期）https://book.yunzhan365.com/tkgd/sash/mobile/index.html

[4] 全球云上数据泄露风险分析简报（第二期）https://book.yunzhan365.com/tkgd/bxgy/mobile/index.html

[5] 全球云上数据泄露风险分析简报（第三期）https://book.yunzhan365.com/tkgd/xyih/mobile/index.html

[6] 全球云上数据泄露风险分析简报（第四期）https://book.yunzhan365.com/tkgd/xbin/mobile/index.html

[7] 全球云上数据泄露风险分析简报（第五期）https://book.yunzhan365.com/bookcase/wxjf/index.html

[8] 云上 LLM 数据泄露风险研究系列（一）：基于向量数据库的攻击面分析 https://mp.weixin.qq.com/s/5jndWjm_yMEXY0E-W369NQ

[9] 云上 LLM 数据泄露风险研究系列（二）：基于向量数据库的攻击面分析 https://mp.weixin.qq.com/s/KZsGvmyE6WtspDb5ZvNKVg

[10] 云上 LLM 数据泄露风险研究系列（三）：开源大模型应用的攻击面分析。

https://mp.weixin.qq.com/s/ADHC4e03ymaPe5ifZ7aODA

大模型安全风险分析与防护架构

绿盟科技 总工办 张睿

摘要：大模型作为新兴 IT 应用技术，因其广阔的通用、专用业务场景，以及高效的智能分析、推理、生成能力，受到各行业积极应用和推广。伴随其应用业态的不断丰富，大模型技术预期影响范围持续扩大。为保障大模型安全、合规使用，企业组织必须以全面的风险管控框架进行风险分析及安全设计，从顶层进行规划，以保证大模型技术在风险受控的前提下导入和应用。

关键词：大模型安全 风险分析 安全合规 防护架构

2025 年 2 月 19 日，国务院国资委召开中央企业“AI+”专项行动深化部署会，强调大模型构建需加速追赶，推动人工智能关键领域取得系列积极进展。当前，以国资企业为代表的各类组织机构迅速部署大模型，研究本领域业务与大模型的融合应用。受国家政策、激励、市场发展前景多重因素驱动，以大模型为主题的各类技术、应用、产品呈现出百花齐放的发展态势，获得了极其广泛的拓展和行业应用。

安全作为大模型技术应用及发展的前提，必须从顶层进行全面规划，在风险受控的基础上导入并应用大模型，从而支撑业务的健康有序发展。大模型安全风险从内容上可以总体划分为安全合规风险、安全技术风险两大类别。基于两类风险，对应设计安全防护架构，可区分独立部署大模型、外联商业大模型两类关键场景，前者从安全基础、安全技术、安全管理三个方面进行综合设计，后者关注数据泄露防护。

1. 大模型安全合规风险

大模型安全合规风险聚焦国家、行业监管机构针对人工智能、大模型应用发布的规章制度及规范性文件，并将相关要求于管理和技术两个层面落实，防止因违规而引发行政处罚等相关风险。对大模型应用法律法规及规范标准梳理过程中，可从两个方向开展。纵向可以从法律法规效力层级逐层分析，以保证全面覆盖相关要求；横向关注区域、行业性要求，合规分析框架如图 1 所示。

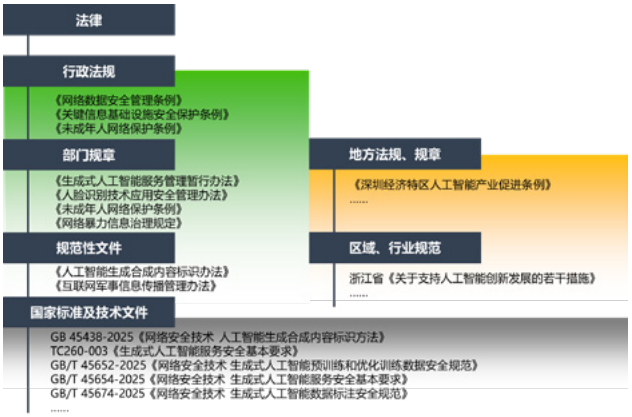


图 1 大模型安全合规风险分析框架

法律及行政法规层面，当前还未发布直接以人工智能、大模型应用为主题的相关文件，但存在涉及生成式人工智能数据安全的条款。如 2025 年 1 月 1 日施行的《网络数据安全管理条例》第十九条规定，对于提供生成式人工智能服务的网络数据处理者，应当加强对训练数据和训练数据处理活动的安全管理，采取有效措施防范和处置网络数据安全风险。以及第四十条，规定了智能终端等设备生产者有关预装应用程序的关联义务 [1]。企业组织机构还需从场景及应用特殊人群方面分析，如前者分析是否涉及关键信息基础设施，以满足《关键信息基础设施安全保护条例》[2]，后者分析是否涉及未成年人，以满足《未成年人网络保护条例》[3]。

部门规章层面，首先必须满足 2023 年 8 月 15 日施行的《生成式人工智能服务管理暂行办法》要求 [4]，其次对于涉及人脸图像生成类的大模型，还需单独进行合规分析，确定满足 2025 年 6 月 1 日起施行的《人脸识别技术应用安全管理办法》要求 [5]。模型内容生成还应对暴力内容进行筛选过滤，以满足《网络暴力信息治理规定》第十二条规定 [6]。涉及特殊行业的，如气象领域，还应进一步满足《人工智能气象应用服务办法》的规定 [7]。

规范性文件、标准及技术文件当前相关内容较多，而且随着时间推移，该部分的内容会持续充实细化，所以需要持续跟进。针对规范性文件，组织机构需要满足国家互联网信息办公室联合其他三部委发布的《人工智能生成合成内容标识办法》[8]，向下关联应用强制性国标《网络安全技术 人工智能生成合成内容标识方法》(GB 45438-2025) 以及技术性文件《生成式人工智能服务安全基本要求》(TC260-003)，并选择适用诸如《网络安全技术 生成式人工智能数据标注安全规范》(GB/T 45674-2025) 等多项推荐性国家标准 [9-11]。

横向有关区域、行业性规范，当前针对人工智能及大模型技术，均以鼓励发展类文件为主，关系到企业组织机构申报政府奖励、补贴、奖项等工作。以合规为主题，依然需要向上引用国家、部门规章类要求，或关联引用本区域、本行业既往有关网络和数据安全要求。

2. 大模型安全技术风险

大模型安全技术风险，需从两类应用场景分别分析，然后进行综合规划。两类场景分别为独立部署大模型场景，即企业组织机构自行部署大模型，能够控制模型的训练、部署、应用的部分或全部阶段；以及外联商业大模型场景，即企业组织机构通过互联网访问开放的商业模型，对模型不具有控制权，如图 2 所示。

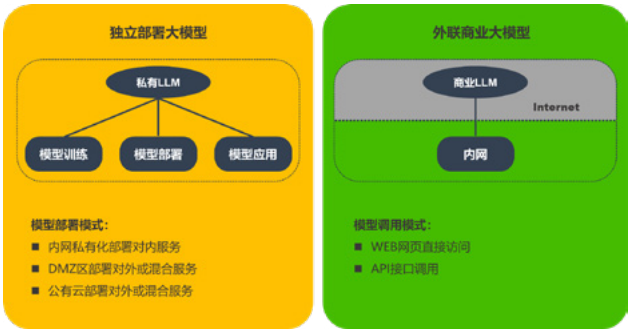


图 2 大模型两类应用场景

独立部署大模型场景，可从时间、空间两个维度进行分析。时间维度从模型训练、部署、应用三阶段划分；空间维度从基座安全、数据安全、模型安全、应用安全、身份安全五个关键领域划分，进行风险矩阵映射（如图 3 所示）。组织机构通过该风险矩阵，能够全面考虑大模型风险，既能支持前期项目可行性研究风险分析，也能于运营期内对风险识别、登记、监控、处置提供指导。该矩阵也便于组织机构基于大模型的生命周期，从训练、部署、应用分阶段梳理安全风险，分角色落实安全责任，从而进行安全设计。以模型应用阶段为例，对外发布大模型服务，技术层面满足五个类别安全要求，由安全部门主导相关安全验证与测试，尤其是上线发布前的综合测试；合规层面整合大模型合规备案、算法备案内容，由业务部门主导，联合安全部门进行备案前安全自查、测试、报告等内容。与此同时，以该风险矩阵为基础，机构也可扩展第三维度，进行能力成熟度评估，以关联机构 IT 能力的持续管理，或可导入《人工智能 大模型 第 3 部分:服务能力成熟度评估》(GB/T 45288.3-2025)，实现本机构的大模型服务能力成熟度自评或未来第三方认证 [12]。

图 3 大模型安全技术风险矩阵

外联商业大模型场景技术风险，主要聚焦数据安全、模型幻觉两个类别，尤以内部数据泄露风险为核心。因互联网商业大模型发展异常迅速，模型互联能力、应用场景丰富程度持续刷新，企业组织机构外联商业模型并提升自身业务能力、处理效率需求强烈，所以该场景一般需由安全部门进行总体设计，并区分 API 外联、WEB 外联两类调用模式，尽快部署安全策略，防止组织内部商业

敏感数据、个人信息乃至涉密数据的外发上传。尤其是该场景下的数据和文件外发风险高于传统网盘、论坛、社交软件，具有暴露面大、影响持久性长的特点。

3. 大模型防护架构

以大模型合规框架、技术风险矩阵为基础，大模型安全防护可以从三个层次展开，如图 4 所示。底层聚焦大模型部署环境，需保障基础设施安全，包括通信网络、区域边界、计算环境、云、容器涉及的安全设计与实现；中层需于技术上实现三类关键业务安全场景，即供应链安全、数据安全和运营安全，于管理上实现大模型合规评估、风险管理、安全监测预警和安全应急响应；顶层实现基座、模型、数据与算法、运行的安全技术目标，以及模型风险可控、合法合规的管理目标。

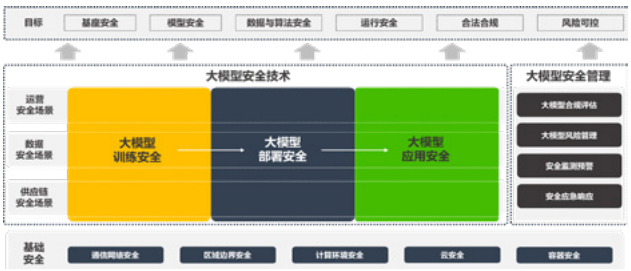


图 4 大模型安全防护架构

大模型供应链安全场景可以结合基础安全进行安全设计，首先是大模型部署网络与应用系统环境安全检测与加固，基于大模型部署环境，需对相关网络、操作系统、云和容器环境进行安全监测与防护；其次，大模型因依赖开源组件，需围绕大模型开源组件的安全检测、许可依赖、脆弱性继承关系进行有效安全防护，并将该能力纳入开发安全流程中，保证后续模型二次训练、微调过程中的代码安全;最后,基于大模型的生命周期,涉及模型开发环境、训练环境、运行环境的安全检测与防护，防止因为三类环境被植入恶意代码、投毒，而导致模型构建、压缩、微调与后续应用期间，出现严重漏洞、数据泄露、违规输出等事件。

大模型数据安全场景必须从两个方面进行安全设计，首先为大模型训练、微调、应用过程中,输入至大模型的数据、文件风险检测，相关数据和文件需进行敏感数据、商业涉知识产权数据、用户隐私信息的识别与风险评估，未加保护可导致数据和商业秘密泄露、逆向工程还原、个人数据违规等风险；其次为大模型输出内容合规过率，因大模型可能被滥用生成违规和虚假信息、恶意代码、仇恨言论等内容，需对大模型输出的内容进行审核，可利用关键词过滤、提示词分类、语义识别等方式，检测并控制大模型生成有害内容的风险，保证符合组织机构安全管理规定、法律法规的要求。

大模型运营安全场景可从两方面设计：技术层面，安全运营需解决大模型运行期间的安全性、可用性问题，安全性涉及智能体调用安全、集成应用或数据源 API 接口安全、模型防越狱及防越权、提示词与思维链注入防护等内容，以及围绕不同层级的身份识别与授权，均为该阶段需要进行详细安全设计和防护的内容，可用性涉及大模型资源滥用和算力耗尽攻击防护，同时需要兼顾网络与操作系统层面 DoS 攻击防护；管理层面，必须完成大模型合规备案，涉及算法合规监管的，还需考虑算法合规备案流程。在此基础上，同步建设大模型风险评估、安全监测预警、应急响应能力，进而整合至组织机构信息安全管理体系，实现统一管理。

4. 结语

大模型技术因其广阔的发展前景受到世界各国的关注和投入，同时也因其不断融合智能体、智能应用相关技术，持续渗透行业各类业务应用，势必带来互联网产业的重大革新，助推传统行业数字化转型，并将加速芯片设计与制造、算力平台构建与应用的研发和升级。安全作为发展的根本，需要统筹协调二者的关系。以全面视角审视大模型安全风险，合理规划大模型应用，以应对合规要求和技术风险，成为企业组织机构应用并推广大模型技术的科学方法。以安全为基，也是行业探索大模型应用新场景，发展大模型技术新业态的可行路径。

参考文献：

[1]《网络数据安全管理条例》

[2]《关键信息基础设施安全保护条例》

[3]《未成年人网络保护条例》

[4]《生成式人工智能服务管理暂行办法》

[5]《人脸识别技术应用安全管理办法》

[6]《网络暴力信息治理规定》

[7]《人工智能气象应用服务办法》

[8]《人工智能生成合成内容标识办法》

[9]GB 45438-2025《网络安全技术 人工智能生成合成内容标识方法》

[10] TC260-003《生成式人工智能服务安全基本要求》

[11] GB/T 45674-2025《网络安全技术 生成式人工智能数据标注安全规范》

[12] GB/T 45288.3-2025《人工智能 大模型 第 3 部分：服务能力成熟度评估》

CAASM+AI+SOAR：重新定义网络资产安全管理

绿盟科技 创新研究院 桑鸿庆、总裁办 张皓天

摘要：在网络安全防护中，资产是攻防双方争夺的核心阵地。本文围绕 CAASM（网络资产攻击面管理）理念，探讨如何通过 AI 与 SOAR 技术提升网络资产的可视性、可控性和响应效率。首先，从边界、内网、泛云与供应链四大类型划定资产范围，明确各类资产的风险特性。其次，通过分析“三无七边”等典型资产风险场景，展示传统资产管理面临的挑战。再系统介绍 CAASM 的核心能力：全域资产发现、智能识别、自动化编排与持续监控，强调 AI 在资产画像构建中的优势，以及 SOAR 对资产管理流程自动化的推动。最后指出，在大模型时代，将 CAASM 与 AI、SOAR 结合，可构建动态化、全生命周期的资产安全管理体系，有效支撑企业“主动防御”战略转型。

关键词：CAASM AI 安全编排（SOAR）攻击面管理

网络安全对抗中，资产是攻防双方争夺的阵地。资产不仅包括硬件设备如服务器、路由器，还涵盖软件系统、数据资源以及网络基础设施等。攻击面是指所有可能被攻击者利用的可访问点，包括外部、内部所有的资产脆弱点。本文将从网络资产范围和资产风险场景入手，介绍 CAASM（Cyber Asset Attack Surface Management）网络资产攻击面管理的思路和功能，以及 AI 和 SOAR 在资产管理中的应用。

1. 网络资产范围

为了有效管理这些资产并减少攻击面，我们首先需要对资产进行分类和界定。一般来说，网络资产按照所属的网域可以分为以下几个方面：

图 1 网络资产分布示意图

边界资产：DMZ（边界网络）连接内网与外网，包括防火墙、路由器、VPN 网关及公网服务等。这些资产直接暴露于外部，是攻击者的首要目标。作为第一道防线，边界资产不应存储任何

机密数据，因此收敛其暴露面尤为关键。

内网资产：内网资产指位于组织私有网络内部的全部资产，包括服务器、工作站、打印机、交换机等，主要通过主动探测、被动流量、端侧发现。由于不直接面对外部网络，其安全能力建设常被忽视。然而，一旦边界被突破，内网资产的安全性将成为重中之重。

泛云资产：随着云原生应用的普及，企业大量服务托管于公有云。云资产包括虚拟机、数据库、存储桶及 DevOps 组件等。尽管云服务商提供基础安全保障，但组织仍需负责云资产的配置和管理，以防因误配置或权限不当引发的风险。

供应链资产：供应链资产指第三方供应商提供的服务和组件。例如，开发团队为调试便利，将客户真实数据（如姓名、订单号及支付 API 密钥）嵌入测试用例并提交至公开仓库，因权限疏忽导致敏感数据暴露。

2. 网络资产风险场景

传统资产安全管理聚焦漏洞识别与修补，而 CAASM 更关注资产本身的暴露状态、配置问题及业务背景。比如资产运营中经常提的“三无七边”资产，“三无”指的是无人管理、无人使用、无人防护情况的业务 / 网站 / 系统 / 平台，“七边”指的是测试系统、试验平台、退网未离网系统、工程已上线加载业务但未正式交维系统、与合作伙伴共同运营的业务或系统、责任交接不清的系统、处于衰退期的系统。基于此梳理一下资产风险场景，帮助安全团队全面理解潜在威胁：

资产类型	定义	示例	风险
影子资产	未被正式记录或管理的设备和应用，未纳入安全管理体系，通常缺乏防护措施	员工私自接入的无线路由器、未经审批的云服务实例、遗忘的测试服务器	未登记，缺乏监控和防护，易成为攻击入口
僵尸资产	已被废弃但未从网络中移除或清理的资产，可能保留权限或漏洞	已经淘汰的服务、未注销云实例、过期但仍可用域名	未清理，无人维护，长期处于高风险状态
两高一弱资产	存在高危端口、高危服务和弱口令的资产，脆弱性明显。	RDP3389 端口服务器、旧版 Apache、弱口令、未授权	高危端口易扫描、服务含漏洞、口令易破解
变化资产	资产特征如 IP、端口、服务信息发生改变	资产系统上登记备案的是网站服务，但实际运行的是数据库。	可能导致安全监控失效增加被攻击的风险
高敏资产	承载敏感数据或关键业务功能的资产，价值高	GitLab、Confluence、CI/CD 组件服务等	出现 0day，失陷后将会对组织造成严重损失
无主资产	缺乏明确归属或责任人的资产，管理缺失	无人认领服务器、未分配云资源、因人员流动失管设备	无人负责，漏洞响应缓慢治理困难
组件盲点资产	系统依赖的第三方组件未登记，影响范围不明	未识别的 Log4j 漏洞组件的 Web 服务	清单缺失，漏洞应急响应资产维度缺失

3. 网络资产攻击面管理（CAASM）

网络资产攻击面管理(Cyber Asset Attack Surface Management，CAASM）是一种先进的安全管理方法，旨在帮助组织全面识别、监控和管理网络资产的攻击面。Gartner 将其定义为“通过与现有工具的 API 集成，提供统一的资产视图，帮助组织识别所有内部和外部资产，查询整合数据，发现漏洞范围和安全控制差距，并进行修复”的新兴技术。与传统漏洞管理聚焦修补已知漏洞不同，CAASM 更关注资产本身的暴露状态、配置问题和业务背景，提供动态、全生命周期的资产管理能力，以应对复杂风险场景，其工作流程如下图所示：

图 2 网络资产攻击面管理工作流程

CAASM 的核心能力：

3.1 全面资产发现

CAASM 的首要能力是自动发现并清点全部网域的资产，包括终端、服务器、云服务、IoT 设备、应用程序以及第三方供应链暴露的资产。通过对多个资产源的融合，动态实时更新资产台账，整合组织内包括内网、边界、云和供应链的资产到统一的视图。

3.2 资产智能深度识别

在复杂多变的网络环境中，精准识别和管理资产是安全防护的基石。借助 AI 技术，CAASM 实现了从资产识别到分类的全流程自动化，构建全面的资产画像。对所有未知资产进行聚类，减少资

产数据处理，利用大模型指纹自动生成技术，为资产生成唯一标识，提高识别精度。资产自动分类将识别结果映射到标准资产库，实现多维度、多领域的资产视图。

3.3 自动化编排

资产的自动化运营可以利用 SOAR(Security Orchestration, Automation, and Response，安全编排、自动化和响应）的思路实现，通过集成资产收集工具、编写剧本来提升资产运营的效率。实现从资产发现、识别、告警到响应的全流程智能化，为安全运营赋能，用剧本驱动资产管理，让资产风险“动态清零”。

3.4 持续监控与威胁情报整合

CAASM 通过周期性持续运营，确保资产视图的动态性与完整性。与传统静态漏洞评估相比，CAASM 利用自动化工具和实时数据集成，以固定周期动态更新资产台账。并且结合漏洞管理、补丁管理及外部威胁情报，实现资产与风险的实时关联，发现异常资产、未授权访问及漏洞资产。通过持续监控与智能分析，CAASM 不仅提升了资产管理的精准性，还增强了组织的整体安全韧性，从资产维度上助力从被动响应转向主动防御。

4. 总结

资产管理常被视为“昨天”的问题，然而随着技术迭代和云化趋势的加速，其复杂性持续困扰着安全运营团队。在大模型时代，这一难题或许能迎来彻底解决。通过大 / 小模型的智能应用，企业可以实现资产标签的自动化与精准化，让每项资产都被清晰分类并标记“身份”。AI 还能辅助挖掘资产间的依赖关系，从海量数据中提炼出攻击面风险的关键洞察，提高资产运营效率。与此同时，SOAR 技术的引入使资产发现、管理和响应的实现资产管理自动化。通过将 CAASM（网络资产攻击面管理）与 AI 和 SOAR 相结合，企业能够构建一个“可见、可管、可防”的全生命周期资产安全管理体系。这种融合不仅有效收敛了资产攻击面，还推动了“主动防御”策略的实现，让安全运营更高效、更具前瞻性。

欢迎对 CAASM 感兴趣的朋友与我们交流探讨！此外，我们免费提供一次资产梳理服务，有需求的朋友可随时联系我们，共同提升资产安全管理能力。

参考文献：

[1].https://www.balbix.com/resources/gartner-take-on-evolving-caasm-the-future-of-cyber-asset-management/.

[2].https://www.gartner.com/reviews/market/cyber-asset-attack-surface-management.

[3].https://www.paloaltonetworks.tw/cyberpedia/what-is-soar.

[4].https://www.runzero.com/blog/new-era-exposure-management/.

可信数据空间（三）数据流通利用设施中的几条路线

绿盟科技 创新研究院 顾奇

摘要：本文系统介绍了《国家数据基础设施建设指引》提出的数据流通利用设施，重点解析“数场”“数据元件”“数联网”三条技术路线的架构理念与实践路径。通过规范化数据抽象、分布式管理和可信计算，这些方案正逐步支撑可信数据空间建设，推动数据要素高效流通与价值释放。

关键词：可信数据空间 数据流通利用设施 数场 数据元件 数联网

一 . 数据流通利用设施

2025 年初，国家发展改革委员会、国家数据局等三部门正式发布了《国家数据基础设施建设指引》[1]。该文件从释放数据要素价值的角度出发，明确提出国家数据基础设施的概念，旨在面向社会提供涵盖数据采集、汇聚、传输、加工、流通、利用、运营、安全等全生命周期服务的一类新型基础设施。

如图 1 所示，国家数据基础设施的构建并非凭空而来，而是基于长期积累的网络设施、算力设施、应用设施等数字基础设施之上进一步拓展与深化。在此框架下，数据流通利用设施被提炼出来，成为数据价值释放的关键组成部分。其中，可信数据空间因其清晰易懂的概念、广泛的社会共识以及较强的实践可操作性，在国家政策的引导下迅速受到各界关注，形成建设热潮。

值得注意的是，在《建设指引》中，数据流通利用设施的技术路线并不止于可信数据空间，而其他路线在实践中往往以组件或理念的方式融入可信数据空间建设方案之中，因而知名度相对较低。本文将在现有公开政策与技术资料的基础上，重点介绍数场、数据元件、数联网三条路线，而对隐私保护计算与区块链等技术相对具体且已有较多介绍性文章，本文不再赘述。

图 1：数据基础设施及网络、算力设施总体架构图

二 . 数场

相较于数据流通利用设施中的其他技术名称，“数场”更偏向于一种规范与框架设计。其核心思想是通过构建标准化技术体系，使数据流通的各个环节更加有序、高效、安全。如图 2 所示，数场的技术架构由五个核心维度组成：

点：点是数据进入数场的入口，它不仅是数据供给方接入的通道，也是数据质量、合规性和安全性的第一道防线。在这个环节，数据需要经过格式校验、资质与合法性审查以及必要的脱敏处理，确保数据符合数场的标准，并防止敏感信息的泄露。

线：线是数场内部的数据传输网络，负责连接各个数据主体和功能平台，确保数据能够在不同节点之间高效、安全、稳定地流动。这一传输网络采用高速光纤、分布式网络架构等技术，确保数据能够以低延迟、高吞吐的方式传输，满足不同场景的需求。

面：面是数场中数据主体、数据资源和计算能力的综合交互空间，是数据流通、共享和交易的核心区域。在这里，数据供需双方可以自由匹配，数据可以在不同主体之间进行共享、交易或联合计算，实现数据价值的最大化。

场：场是基于数场基础设施构建的数据应用和创新生态，它是数据从静态资源转化为实际价值的关键环节。数场不仅提供数据存储和流通能力，还构建了丰富的行业应用生态，使数据能够在金融、医疗、交通、工业等多个领域发挥作用。

安全：安全是数场的核心保障体系，覆盖数据接入、传输、存储、计算和交易的全生命周期，确保数据在整个流通过程中的安全性与合规性。数场采用多层次的安全策略，包括数据加密、隐私保护计算、访问控制、数据溯源等机制，构建全方位的安全体系。

图 2：数场功能架构图

三 . 数据元件

数据元件由中国电子信息产业集团提出并推广，如图 3 所示，其核心思想是通过对数据的合理抽象，使数据成为稳定的要素形态，从而在保障安全性的同时，促进数据的高效流通与价值释放，并进一步支撑了数据要素化的流通模型与安全模型。

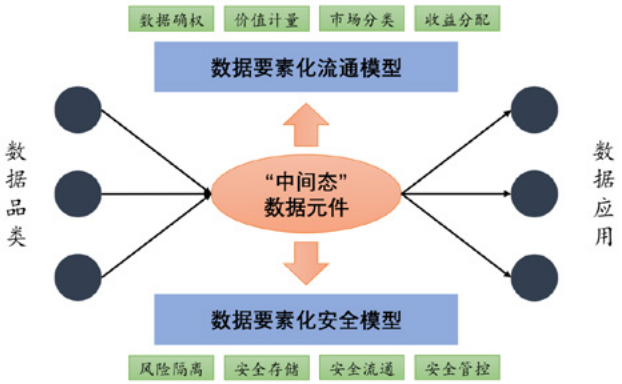


图 3：基于数据元件的数据要素流通

3.1 数据抽象的三个阶段

数据元件的提出，实际上承载了数据抽象演进的历史脉络，主要经历了以下三个阶段：

数据与应用程序的解耦：在传统计算模式下，数据与应用程序高度耦合，导致系统开发复杂度较高。数据库系统的出现，通过数据表结构的标准化抽象，使结构化数据能够独立于应用程序，从而降低了软件开发的门槛。

数据与业务系统的解耦：随着企业数字化转型的深入，业务应用的多样性与数据形态的复杂性日益增长，传统的企业级数据管理模式已难以满足需求。因此，数据湖、湖仓一体等架构应运而生，进一步屏蔽了企业内部数据汇聚与分析的复杂性，推动了数据的跨业务系统共享与复用。

数据与社会主体的解耦：进入数据要素化时代，数据的价值

释放已不仅限于企业内部，而是需要在更广泛的社会范围内实现流通与利用。数据元件正是基于这一需求提出，通过标准化的数据抽象方式，使数据能够在不同主体之间安全、高效地流通。

3.2 数据件

在数据元件的基础上，孙凝晖院士进一步提出了“数据件”的概念，并明确了数据件应满足的四大核心要求：

可寻址：数据件需具备唯一标识与寻址机制，确保在广域范围内可被精准定位与访问。

可交换:数据件在不同主体、不同系统间应具备语义互通能力，确保数据流通的准确性与一致性。

可操作：数据件应提供标准化的访问与操作接口，使其可以即插即用，并支持进一步深度加工。

可管控：数据件需具备内生的安全管控机制，确保数据在流通过程中的合规性与安全性。

基于上述四个要求，如图 4 所示，孙院士进一步提出了数据件基本结构，对此有兴趣的读者不妨进一步阅读 [2]。数据件的提出，进一步深化了数据标准化抽象的实践路径，为可信数据空间的建设奠定了坚实的技术基础。

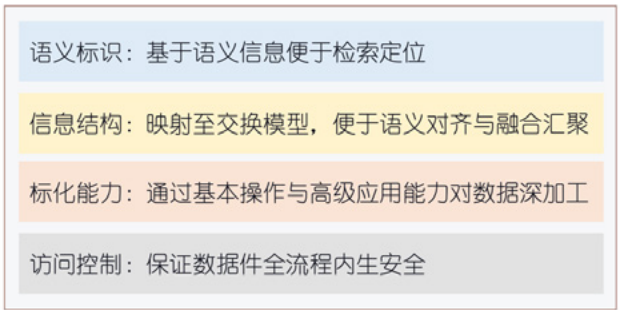


图 4：数据件基本结构

四 . 数联网

4.1 BDWare 开源方案

数联网概念最早由北京大学黄罡教授团队提出，旨在解决互联网环境下数据分散、低效访问、难以复用等问题。数联网提出了一种新的数据组织方式，通过数据一阶实体化，使数据成为直接可用、可操作、可寻址的独立逻辑实体，从而构建一个真正以数据为中心的网络空间。在这一架构下，数据本身不再依赖于某个特定的物理存储位置，而是通过分布式方式进行管理、调度和流通，极大提高了数据的可用性和共享价值。该方案也形成了 BDWare 开源软件 [3]。

如图 5 所示，在该方案中，其结合数字对象架构（DOA）、智能合约、分布式账本等技术，致力于形成一套完整的数据空间基础设施，包括如下核心部分：

基于数字对象架构的一阶数据实体模型及交互技术：采用数字对象架构（DOA），将数据拆分成标识、元数据、实体三部分，使数据可以被唯一标识、有效存储和灵活管理;通过扩展的 DOIP（Digital Object Interface Protocol）解决数据访问的网络环境依赖问题，使数据访问不再依赖于特定的通信协议（如 HTTP 或 TLS）。

基于语用合约的一阶数据实体使用技术：传统互联网数据使用方式是数据提供方决定数据的使用方式，而数联网引入语用合约，让数据使用方式由数据需求方定义，并与数据提供方达成共识。语用合约类似于智能合约，确保数据的使用方式透明可控，同时保障供需双方的权益。

基于关系链的一阶数据实体可信保障技术：由于数据在数联网中是自由流通的，数据的使用关系会变得非常复杂。为此，数联网采用关系链系统，基于区块链等分布式账本技术，为每次数据操作建立可信溯源机制。通过分层随机共识技术,提高区块链的吞吐量，

使其能够支持互联网级别的数据关系记录。

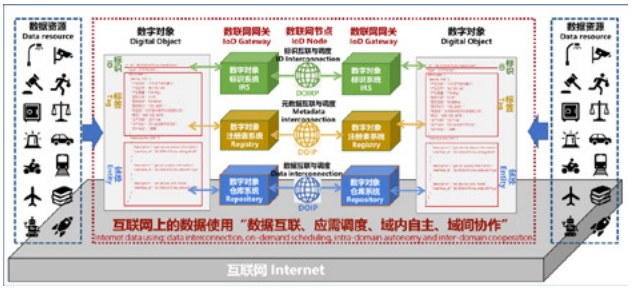


图 5：基于数字对象架构的数联网基础设施

4.2 中国移动 DSSN 方案

无独有偶，中国移动在 2023 年也提出了一种数联网架构 [4]，其理念更接近于计算机领域经典的“加中间层软件”思路，采用“交易 - 交付分离 + 网络化可信计算”的架构，通过标准化的数据流通协议、可信计算技术、数据资产管理体系，构建一个跨行业、跨区域、跨主体的数据流通网络。其具体包含：

DSP（数据交付平台）：负责数据交易管理、任务调度、交易撮合、流通全链路管控等；

DSN（数据服务节点）：部署在数据提供方，实现数据源对接、DSSN 专网接入、数据安全计算等功能；

DRN（数据需求节点）：部署在数据需求方，实现业务系统对接、DSSN 专网接入，数据可视化开发、数据计算等功能；

这些网元之间通过 DSSN 专网连接，形成一个分布式、可扩展的可信数据流通网络，使数据可以在不同主体之间安全流通。笔者以为，相较于其他数据流通方案，DSSN 更强调充分利用现有技术体系，以多维度分层适配数据要素流通中的关键挑战，提供更具可行性和可扩展性的解决方案。

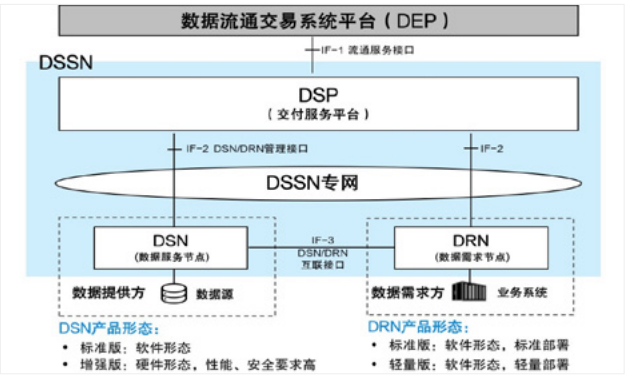


图 6：中国移动数联网方案

五 . 总结

本篇文章介绍了数据流通利用设施中的几条重要路线，不难看出，这些路线既代表了不同角度的技术探索，也在实践中不断融合借鉴与创新，是国内多类数据要素主体和市场发展脉络的真实体现。

展望未来，可信数据空间的建设将以国际数据空间中的使用控制理念为基础，吸收多元理念与前沿技术，构建更加成熟、完善的体系，助力数据要素市场的高效流通与价值释放。在下一篇文章中，我们将结合信通院等权威机构的指导方案，以功能视角深入剖析可信数据空间的建设方法与核心组件。

参考文献：

[1]https://www.gov.cn/zhengce/zhengceku/202501/P020250106393009877184.pdf

[2] 孙凝晖 , 郭嘉丰 . 数据件：一种数据要素标准化抽象 [J]. 中国计算机学会通讯 , 2024, 20(10): 1-10

[3]https://gitee.com/BDWare

[4] 中国移动研究院用户与市场研究所 . 面向数据要素流通的新型基础设施——数联网（DSSN）[R]. 北京 : 中移智库 , 2023.

网盘数据泄露探索：从访问控制突破到敏感信息发现

绿盟科技 创新研究院 浦明

摘要：随着云存储和网盘服务的广泛普及，网盘数据泄露问题日益突出，已成为影响企业和个人信息安全的重要隐患。本文围绕网盘数据泄露展开研究，重点探讨访问控制机制中的安全漏洞及敏感信息发现方法，结合近年来典型案例与实际调研，揭示网盘分享链接在安全配置上的普遍风险。通过分析多种信息收集手段，包括测绘引擎和代码仓库等，揭示大量敏感文件因访问权限配置不当而暴露在公共网络环境中，进一步说明网盘数据泄露的广泛性和严重性。最后，本文分享绿盟科技星云实验室在云上风险发现领域的创新研究方案，旨在为网盘安全防护提供切实可行的技术思路和实践指导。

关键词：数据泄露 标签 网盘 威胁情报

1. 概述

数据泄露已成为数字时代企业和个人面临的重大威胁之一，不仅会造成经济损失，还可能损害企业声誉和用户信任。因此，深入分析数据泄露的趋势和成因至关重要。

2024 年，国际知名咨询机构 Verizon 发布的《2024 年数据泄露调查报告》（2024 Data Breach Investigations Report）[1] 指出，“68% 的数据泄露事件涉及人为错误，如配置错误、社工攻击、话术欺诈等，人依然是安全链条上的重要一环”[2]。这一结论与近年来频发的云安全事件 [3,4,5,6,7] 相呼应，表明由“人为因素”引发的数据泄露问题将持续存在。

绿盟科技创新研究院在云上风险发现和数据泄露领域深耕多年，并发布了多篇相关研究报告 [3,4,5,6,7]。早在 2021 年，研究院就针对对象存储导致的数据泄露问题进行了深入研究 [8]。然而，数据泄露问题不仅限于对象存储，网盘数据外泄事件也屡见不鲜。以下是几起典型的历史事件 [13]：

2014 年 iCloud 泄密事件：攻击者通过钓鱼攻击和暴力破解等手段，获取了苹果公司 iCloud 的访问权限，盗取了大量名人的私密照片和视频。这些内容随后被公开，引发了广泛的社会关注。

2015 年百度网盘数据泄露事件：攻击者利用百度网盘的系统漏洞，获取了部分用户的个人文件和隐私信息，并将其公开在互联网上。

2016 年 Dropbox 数据泄露事件：Dropbox 披露其在 2012 年遭遇的数据泄露事件中，有 6800 万个用户账号信息被盗，包括邮箱地址和加密后的密码。事件的根源是 2012 年的漏洞未及时修补，导致数据在 4 年后被发现外泄。

2023 年 12 月日本 Ateam Entertainment 数据泄露事件：日本游戏开发商 Ateam Entertainment 母公司公告称，近百万人的个人信息因 Google Drive 配置错误泄露近 6 年。

2024 年 9 月阿里网盘数据泄露事件：阿里网盘因系统 BUG 导致用户个人隐私照片泄露 [9]。多名用户反映，在阿里网盘 PC 端相册中创建新文件夹时，系统会自动加载陌生人的隐私照片和视频，这些内容不仅可预览，还可直接打开。

此外，2023 年底，数据安全公司 Metomic 的一份报告显示，存储在 Google Drive 上的文件中，有 40.2% 包含敏感数据。Metomic 分析了约 650 万个 Google Drive 文件，发现 34.2% 的文件与公司域外的外部联系人共享，超过 35 万个文件可供公开访问，任何拥有链接的人均可不受限制地查看。

基于上述历史事件和报告分析，笔者近期在网盘安全领域展开了相关研究。本文将从以下几个方面展开：首先，介绍主流网盘的访问控制机制及其潜在安全隐患；其次，探讨针对网盘信息收集的主要方法，包括测绘引擎、代码仓库、Google Hacking 等，并分析网盘内敏感信息的发现途径；最后，分享绿盟科技创新研究院在云上风险发现领域的创新研究方案，以期为读者在网盘泄露防护方面提供新的思路与启发。

2．网盘文件访问控制机制探索

当今云存储市场中，国内外网盘厂商众多，各具特色。国外以 Google Drive、Microsoft OneDrive 和 iCloud 等大厂推出的网盘产品为主，而国内则以百度网盘、阿里网盘、腾讯微云和 360 网盘等为代表。此外，开源类网盘如 Seafile、Cloudreve 和 Nextcloud 也在市场中占据一席之地。尽管网盘产品种类繁多，但笔者在研究其文件访问控制机制时发现，这些机制在本质上具有较高的相似性。限于篇幅，本文将以国外的 Google Drive 和国内的百度网盘为例进行说明。

2.1 百度网盘文件访问控制机制

百度网盘为用户提供了三种文件或文件夹的访问方式：通过链接访问、通过客户端 /Web 端密码登录方式访问，以及目标文件访问地址结合 Token 及 SDK 形式的访问。本文重点介绍通过链接的访问方式，并对其访问控制机制进行详细分析。

百度网盘通过链接的访问方式具有以下特征：

访问有效期：必填项，用户可选择 1 天、7 天、30 天或永久有效。

公开访问配置：用户可选择是否配置提取码。若配置提取码，则链接为非公开访问，提取码可由系统自动生成或由用户手动设置。

授权对象：链接可分享给任何人或仅限于网盘好友。

图 1 百度网盘访问控制配置页面

从上述特征可以看出，百度网盘的访问控制机制主要依赖于提取码和有效期的结合。此外，分享链接的格式会因是否配置提取码以及用户使用的是个人版还是企业版而有所不同。需要注意的是，百度网盘在生成分享链接时，会对初始链接进行 302 重定向，最终生成实际访问链接。以下是不同场景下的链接格式说明：

个人版分享链接格式

1. 公开分享（无须提取码）

生成链接：https://pan.baidu.com/s/<22 位随机字符 >

实际访问链接：https://pan.baidu.com/share/init?surl=<22 位随机字符 >

2. 有提取码分享

生成链接：https://pan.baidu.com/s/<22 位随机字符 >?pwd=<4 位数字 >

实际访问链接：https://pan.baidu.com/share/init?surl=<22 位随机字符 >&pwd=<4 位数字 >

企业版分享链接格式

1. 公开分享（无须提取码）

生成链接：https://pan.baidu.com/e/<22 位随机字符 >

实际访问链接：与生成链接相同

2. 有提取码分享

实际访问链接：https://pan.baidu.com/e/verify?surl=<22 位随机字符 >（需要输入提取码）

个人版与企业版的区别

个人版和企业版的网盘文件分享链接格式存在一定差异：

个人版：链接前缀为 https://pan.baidu.com/s/<22 位随机字符 >，其中参数为 s/<22 位随机字符 >。

企业版：链接前缀为 https://pan.baidu.com/e/<22 位随机字符 >，其中参数为 e/<22 位随机字符 >。

2.2 Google Drive 文件访问控制机制

Google Drive 实际上也是通过共享链接的方式去查看文件信息的，但访问控制机制略有不同：

访问有效期：个人版无有效期设置，企业和校园版才有有效期设置 [12]。

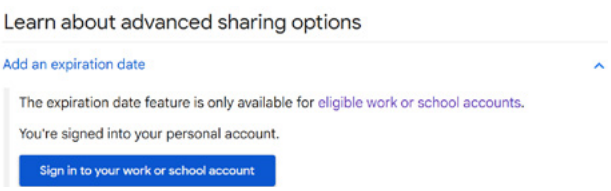


图 2 Google Drive 文件分享有效期设置

- 公开访问配置

Google Drive 中，文件的公开访问配置主要分为两种模式：Restricted（禁止公开）和 Anyone with the link（公开访问）。这两种模式决定了文件或文件夹的访问权限范围。

Restricted（禁止公开）：当选择此模式时，文件或文件夹的访问权限将被严格限制。分享链接时，必须输入被授权访问对象的 Google 账号。授权完成后，可以进一步设置访问对象的具体操作

权限，如图 3、4 所示，包括：

Viewer（可查看）：仅允许查看文件内容，无法进行编辑或评论。

Commenter（可评论）：允许查看文件内容并添加评论，但无法编辑文件。

Editor（可编辑）：允许查看、评论并编辑文件内容。

配置完成后，浏览文件分享链接将自动跳转至 Google 账号的登录页面。只有被授权的 Google 账户才能访问该文件，否则系统会提示“需要申请权限”才能访问，如图 5 所示。

Anyone with the link（公开访问）：当选择此模式时，任何拥有该链接的用户都可以访问文件或文件夹，且无须登录 Google 账号。这种设置适用于需要广泛共享的文件，但也可能带来数据泄露的风险，尤其是在未设置访问限制的情况下。

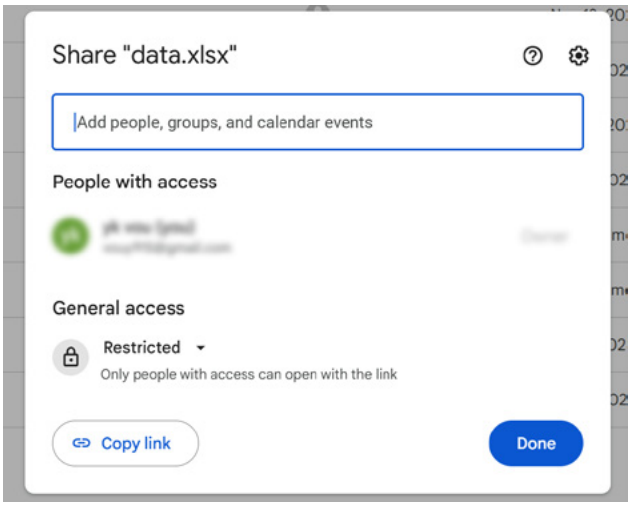


图 3 Google Drive 文件分享链接页面 1

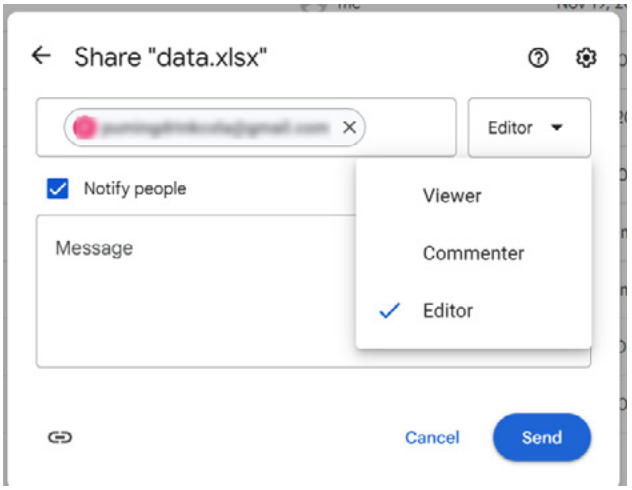


图 4 Google Drive 文件分享链接页面 2

图 5 Google Drive 文件分享链接页面 3（无权限访问文件）

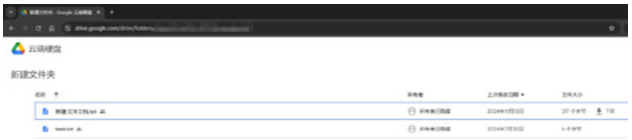


图 6 Google Drive 文件分享链接页面 4（公开访问）

- 访问对象：Google Drive支持多种访问对象类型，包括：

授权 Google 账户：指定具体的 Google 账号，只有这些账号才能访问文件。

账户组：可以将多个 Google 账号组成一个群组，并授权该群组访问文件。

任何人:允许所有拥有链接的用户访问文件，无须登录 Google 账号。

- Google Drive分享链接格式

Google Drive 的分享链接主要分为两种类型：drive 和 docs。具体格式如下：

drive 类型:适用于 Blob 类型的文件或文件夹，生成的链接格式为:

文件：https://drive.google.com/file/d/<33 位随机字符 >

文件夹：https://drive.google.com/drive/folders/<33 位随机字符 >

docs 类型：适用于 Google Workspace 文档，如 Google Docs、Google Sheets、Google Slides 等，生成的链接格式为：

Google Docs：https://docs.google.com/document/d/<44 位随机字符 >

Google Sheets：https://docs.google.com/spreadsheets/d/<44 位随机字符 >

Google Slides：https://docs.google.com/presentation/d/<44 位随机字符 >

根据上述访问控制机制，如果未正确配置访问权限，可能会导致数据泄露，其中：

百度网盘：如果将文件或文件夹设置为公开分享，且未设置提取码或过期时间，则任何人在任何时间都可以访问该文件，存在较大的数据泄露风险。

Google Drive：如果分享文件时未限制特定 Google 用户或群组访问，且设置为 “Anyone with the link”，则任何人都可以在匿名情况下访问该文件，同样存在数据泄露的隐患。

3. 网盘信息收集方法探索

从上述内容不难看出，生成并分享一个公开的网盘链接是非常简单的操作。然而，正是这种便捷性导致了每天都有大量的网盘文件分享链接被生成并传播。如果这些链接未经过严格的访问控制，网盘数据外泄的问题将变得尤为严重。特别是当这些文件中包含敏感信息（如公民隐私数据、商业机密等）时，可能会导致隐私泄露、法律纠纷，甚至面临执法部门的追责。

如前文所述，通过在浏览器中输入正确的网盘分享链接，任何人都可以访问并获取网盘内的文件内容。因此，如何获取这些网盘链接成为一个关键问题。近期，笔者针对多种信息源渠道进行了相应研究与测试，包括测绘引擎、公共代码仓库、自建仓库以及 Google Hacking 等技术手段。测试结果表明，大量已存活的网盘

分享链接被暴露在公网上，进一步证实了网盘数据外泄问题的普遍性和严重性。

注：文中案例相关操作均在实验环境中进行，相关技术仅供研究交流，请勿应用于未授权的渗透测试。

3.1 测绘引擎

开源情报（OSINT）领域，Shodan 和 Fofa 等测绘引擎常被用于查询现网存活的资产信息。本文以这两款主流测绘引擎为例，探讨如何通过指纹（如 icon_hash，title 等）来查询暴露的网盘链接。

图 7 通过 icon_hash 获取网盘资产信息

通过分析查询结果，我们可以从 Title 字段中识别出资产的类型。

图 8 Fofa Title 网盘统计

其中，占比最高的 Title 为“百度网盘 |Synology”，笔者推测这些资产可能与群晖科技与百度网盘合作的增值套餐有关 [11]。值得注意的是，这些资产大多为私有地址，进一步验证了其可能的企业或特定用户属性。由于其他 Title 对应的资产数量较少，笔者暂未对其进行深入分析。

类似地，Google Drive 的分享链接也可以通过特定的指纹进行定位，例如使用“Location:https://drive.google.com/drive/folders/”作为查询条件。

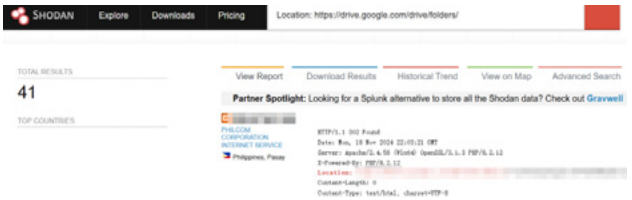


图 9 shodan 查询 google drive 页面

虽然通过开源情报工具可以获取部分网盘分享链接，但由于时效性和数据更新频率的限制，这种方法在实际应用中可能并不高效。未来研究可以探索结合其他技术手段(如动态爬虫或 API 接口)来提高获取有效链接的成功率。

3.2 代码仓库

笔者也尝试了通过代码仓库获取网盘链接，这里我们主要介绍公共代码仓库爬虫和自建代码仓库发现两种方式。

3.2.1 公共代码仓库爬虫

项目开发过程中，开发者可能会无意中将网盘分享链接硬编码到代码中，并将代码上传至公共代码仓库 (如 Github、Gitlab、Gitee 等)，导致这些链接长期暴露在互联网中。针对这一问题，可以通过调用公共代码仓库的官方 restful API，结合网盘路径特征 (如第二节中介绍的分享链接格式) 进行内容检索，并利用正则表达式提取网盘分享链接。具体而言，可以针对 Github 的 Code、Commits、Issues、Repositories 等多个维度进行 API 检索。然而，由于不同公共代码仓库的检索机制存在差异，返回的结果数量也会有所不同。

图 10 公共代码仓库调用不同维度 API 返回的结果数

经过测试发现，通过公共代码仓库获取的网盘分享链接数量远高于通过测绘引擎得到的数量。这一现象间接反映了开发者将硬编码的网盘链接上传至公共代码仓库的普遍性和严重性。

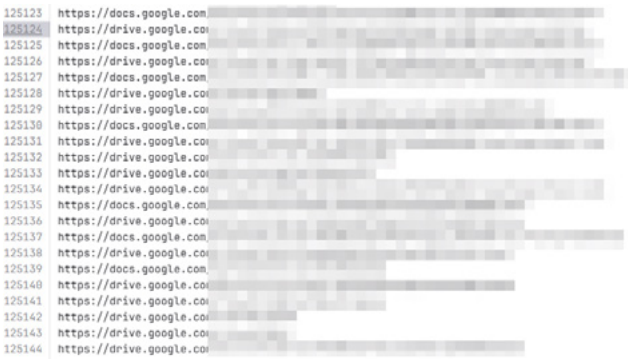


图 11 笔者测试公开的网盘分享链接

3.2.2 自建代码仓库发现

由于笔者所在团队从事云上风险发现已有多年经验，针对已存在泄露风险的自建源码仓库中我们也进行了测试（已得到监管机构相关授权），结论也是发现了大量网盘分享链接，且数量很多。

图 12 已泄露的自建源码仓库文件中发现云盘链接截图

图 13 笔者测试已泄露的自建源码仓库文件中发现云盘链接

3.2.3 Google Hacking

Google Hacking 是一种利用搜索引擎的高级搜索语法来识别 Web 应用程序安全漏洞、收集目标信息、发现泄露敏感数据的错误消息以及定位包含凭据或其他敏感文件的强大技术。笔者通过这一技术，成功从公网中发现了大量暴露的网盘分享链接。

以百度网盘为例，通过使用“site:https://pan.baidu.com/s/”的搜索语法，可以获取到以“https://pan.baidu.com/s/”为前缀的URL，从而定位暴露在公网上的百度网盘分享链接。随后，通过爬虫技术对这些链接进行批量收集和整理，能够显著提高信息收集的效率。

图 14 duckduckgo 发现的百度云盘链接

在获取到网盘分享链接后，下一步需要验证这些链接是否为公开可访问的网盘资源。笔者通过分析具体页面的响应内容，结合网页的 Title 以及请求响应中的特征信息（如状态码、页面结构等），对链接的有效性进行了筛选和验证。经过测试，成功获取了大量公开的网盘访问链接。

```
462    "https://pan.baidu.com/s/[redacted]: [
463        {
464            "type": "[redacted]
465            "name": "[redacted]
466            "date": "[redacted]
467            "size": "[redacted]
468            "expirat[redacted]
469        }
470    ],
471    "https://pan.baidu.com/s/[redacted]  [
472        {
473            "type":
474            "name":
475            "date":
476            "size":
477            "expirat
478        },
479        {
480            "type": "
481            "name": "
482            "date": "
483            "size": "
484            "expirati
485        },
486        {
487            "type": "
488            "name": "
489            "date": "
490            "size": "
491            "expirati
492        }
493    ],
494    "https://pan.baidu.com/s/[redacted]  [
495        {
496            "type": "dir",
```


图 15 通过爬虫测试发现的网盘链接及内容

4. 网盘内敏感信息发现

正常情况下，网盘分享链接被配置为公开可访问前提是所有者期望对方能去访问这些信息且是不包含敏感信息的，但经笔者测试发现，事实上仍然存在较多的敏感信息被泄露，如 2024 年 11 月，由绿盟科技创新研究院发现美国一所中小学使用的 Google Drive 存在访问配置错误，导致该校的审计财务报表、预算报表、支出报表、教师雇佣合同等敏感信息泄露 [7]。

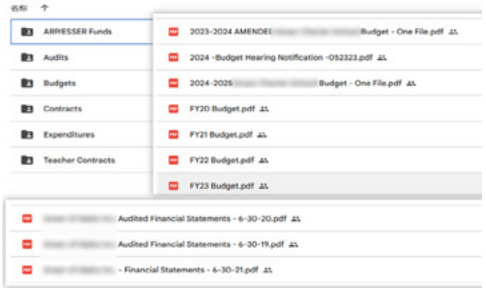


图 16 疑似 Google Drive 隐私信息泄露截图

图 17 疑似印度公民隐私信息泄露截图

图 18 Fusion 能力全景图

5. 总结

鉴于以上分析内容，人为配置错误导致的网盘数据泄露事件是真实存在并时刻在发生，作为安全从业人员，建议相应的防护策略从两方面出发，一方面加强安全意识，非必要情况下尽量在分享链接时避免设置公开访问选项，且有效时长给予较短时间；另一方面使用云上风险监测服务，当发生网盘数据泄露告警时，及时进行权限配置更改避免敏感数据外泄。

Fusion 是由绿盟科技创新研究院研发的一款面向数据泄露测绘的创新产品，集探测、识别、泄露数据侦察于一体，针对互联网中暴露的泛云组件进行测绘，识别组件关联的组织机构和组件风险的影响面，实现自动化的资产探测、风险发现、泄露数据分析、责任主体识别、数据泄露侦察全生命周期流程。

Fusion 的云上风险事件发现组件具有如下主要特色能力：

资产扫描探测：通过多个分布式节点对目标网段 / 资产进行分布式扫描探测，同时获取外部平台相关资产进行融合，利用本地指纹知识库标记，实现目标区域云上资产探测与指纹标记。

资产风险发现：通过分布式任务管理机制对目标资产进行静态版本匹配和动态 PoC 验证的方式，实现快速获取目标资产的脆弱性暴露情况。

风险资产组织定位：利用网络资产信息定位其所属地区、行业以及责任主体，进而挖掘主体间存在的隐藏供应链关系及相关风险。

资产泄露数据分析：针对不同组件资产的泄露文件，结合大模型相关技术对泄露数据进行分析与挖掘，实现目标资产的敏感信息获取。

参考文献：

[1] https://www.verizon.com/business/resources/reports/dbir/ .

[2] https://puming.zone/post/2024-08-27-2024-verizon-dbir%E8%A7%A3%E8%AF%BB-%E6%95%B0%E6%8D%AE%E6%B3%84%E9%9C%B2%E8%BD%AC%E5%90%91%E8%BF%9E%E6%8E%A5%E4%BA%91%E7%9A%84%E7%AC%AC%E4%B8%89%E6%96%B9%E8%BD%AF%E4%BB%B6%E4%BE%9B%E5%BA%94%E9%93%BE/.

[3]《2023 公有云安全风险分析报告》https://book.yunzhan365.com/tkgd/qdvx/mobile/index.html .

[4]《2024 上半年全球云上数据泄露风险分析报告》https://book.yunzhan365.com/tkgd/cltc/mobile/index.html .

[5] 全球云上数据泄露风险分析简报（第一期）https://book.yunzhan365.com/tkgd/sash/mobile/index.html .

[6] 全球云上数据泄露风险分析简报（第二期）https://book.yunzhan365.com/tkgd/bxgy/mobile/index.html .

[7] 全球云上数据泄露风险分析简报（第三期）https://book.yunzhan365.com/tkgd/xyih/mobile/index.html.

[8] https://mp.weixin.qq.com/s/N5f9hqg3swg1CFhXTZMtLQ.

[9]https://mp.weixin.qq.com/s?__biz=MzI2NDEzMzY1Mg==&mid=2652578906&idx=2&sn=a564ba108338a5aeb8e937edcbf3cc4c.

[10] https://www.163.com/dy/article/INH2423405128DFG.html.

[11] https://www.synology.com/zh-tw/dsm/packages/baiduapp.

[12] https://support.google.com/drive/answer/2494822?hl=zh-Hans&sjid=15617540479346960914-NA.

[13] https://eyun.baidu.com/content/115233/ .

解读｜首部全国性政务数据共享法规出台

绿盟科技 数据安全BG 王佳

2025 年 6 月，我国首部《政务数据共享条例》正式颁布（以下简称《条例》），自 2025 年 8 月 1 日起实施，标志着政务数据开放共享进入“有法可依”新阶段。《条例》明确要求建立“统一标准、安全可控、高效流通”的数据共享机制，为政务数据跨部门、跨层级、跨区域流通提供了制度保障。绿盟科技结合长期在数据安全领域的研究积累，深入解读《条例》核心要点，并分享在可信数据流通方向上的相关探索与实践。

1.《条例》核心要点解读

1. 1 立法目的与适用范围

- 目的：推进政务数据安全有序高效共享，提升政府数字化治理能力和政务服务效能，全面建设数字政府。
- 适用范围：政府部门及法律、法规授权的组织之间的政务数据共享活动，不包括国家秘密、工作秘密数据。

1.2 管理体制与职责分工

- 统筹机构：国务院政务数据共享主管部门统筹全国工作，县级以上地方主管部门负责本区域统筹。
- 部门职责：政府部门需设立专职机构，负责数据目录编制、共享申请审核、安全性评估等职责。

1.3 目录管理制度

- 统一目录：政务数据实行统一目录管理，按共享属性分为无条件共享、有条件共享、不予共享三类。
- 动态更新：因法规或职责调整需更新目录的，部门需在10个工作日内完成更新并报审。
- 禁止设障：严禁擅自增设条件阻碍共享，不予共享类需列明法律依据。

1.4 共享使用规则

- 禁止重复收集：通过共享可满足履职需要的，不得向公众重复收集数据。
- 时效要求：
  - 无条件共享数据：1个工作日内答复。
  - 有条件共享数据：10个工作日内答复。
- 数据回流：上级部门需向下级及时、完整回流属地数据，不得额外设限。

1.5 安全保障义务

- 责任原则：按照“谁管理谁负责、谁使用谁负责”原则明确安全责任主体。
- 技术措施：要求采取技术手段防止数据篡改、泄露或非法利用。

2. 绿盟科技可信数据空间方案支撑合规落地

绿盟科技积极响应《条例》要求，推出可信数据空间解决方案，助力政务数据安全共享与价值释放。

2.1 可信连接器：数据流通的“安全桥梁”

绿盟科技研发的可信连接器是各数据流通主体互联互通的核

心组件，具备以下能力：

- 统一接入标准：支持多源数据安全接入，确保数据格式与接口的标准化。
- 严格认证机制：基于数字身份认证与动态权限控制，防止未授权访问。
- 全程可追溯：数据流转全过程审计，实现操作留痕与责任追溯。

同时，可信连接器凭借数据资产发现、数据沙箱、数据脱敏与水印等技术，支撑可信连接器的数据流通利用能力，实现了数据的可信共享和高效利用。基于可信根形成集远程证明、可信启动、动态度量于一体的可信启动链，同时内置轻量化安全防护能力，对连接端提供主动免疫的安全保障，构建内生主动安全防御体系。基于多方安全计算、联邦学习、机密计算等隐私计算技术，提供数据安全计算能力。

2.2 可信数据空间方案：全流程安全管控

绿盟科技可信数据空间解决方案以“数据流通全流程可信可控”为目标，深度融合隐私计算、可信执行环境（TEE）等核心技术，确保数据从采集、存储到使用的全程安全：

- 采：通过可信连接器安全采集数据，确保来源真实可信。
- 存：采用密态存储技术，防止数据泄露与篡改。
- 管：基于智能数字合约实现数据使用权限的动态管控。
- 用：通过联邦学习、机密计算等技术，实现数据“可用不可见”。

目前，该方案已在医疗、政务等多个领域成功落地，帮助客户构建安全、合规的数据共享生态。

《条例》的出台，标志着我国数据要素市场化进入新阶段。绿盟科技将持续深耕数据安全领域，以可信连接器和可信数据空间为核心，助力客户构建合规、高效、安全的数据共享生态，共同推进数字中国建设！

网络安全政策导读（2025年3月—6月）

绿盟科技 总体技术部 林涛

栏目说明：

本专栏基于绿盟科技政策研究团队在网络安全政策法规方面的日常跟踪，筛选国内外当期热点政策法规文件，并重点结合网络安全产业发展，对其内容和影响等进行简要分析。本期研究的国内外政策法规的发布时间范围为 2025 年 3 月—6 月。

更多内容敬请关注微信公众号：“绿盟科技”和“网络安全罗盘”。

1. 国内篇

1.1《人工智能生成合成内容标识办法》发布

（一）从内容看，《办法》主要明确了两个主要问题

一是明确了标识的分类。《办法》首次明确标识包括显式、隐式两类。二者的主要区别之一在于能否“被用户明显感知到”；还有一个重要区别在于标识的添加位置和形式的不同：根据《网络安全技术人工智能生成合成内容标识方法》（GB45438-2025），显式标识一般是直接在生成内容或交互场景界面上添加文字、声音或图形标识，而隐式标识则需要在生成内容的元数据中添加特定要素或对内容添加数字水印。

二是明确了管理对象应承担的主要义务。《办法》主要规定了生成合成内容提供者和传播者、分发平台和用户四类主体的主要义务。对于生成合成内容提供者，主要应承担添加标识、服务协议说明和提示、提供无显式标识内容特殊要求、备案评估衔接等 4 项义务；对于生成合成内容传播者，主要应承担核验隐式标识、添加提示标识、提供标识功能等 3 项义务;对于分发平台，主要需承担上线提示、核验等 2 项义务；对于用户，主要应承担传播生成合成内容时的声明和标识义务。

（二）《办法》的思考和产业影响

一是关于数字水印的分类问题。按照对《办法》第五条的字面理解，数字水印被划归为隐式标识之一。而在实践中，数字水印还常常以显式的方式存在于多种形式的内容成果中，其在表现形式上也具有某些显式标识的特征。是否需要将数字水印这种特定的标识进行进一步的细分，期待相关法规标准后续做出统筹安排。

二是对于“删除、隐匿”标识等情形的认定问题。用户对于生成合成内容成果的使用，很多情况下并非整体使用，更多或表现为

对生成合成内容进行节选或筛选使用，在此类情况下，如何规范对标识的使用，或需进一步明确。

三是产业影响。《办法》及相关标准的发布，无疑将会对数据技术和数据安全行业带来一定市场机会。一方面，从技术保障来看，无论是显式标识还是隐式标识，都需要数据标识相关技术和产品的支撑，重点行业涉及标注咨询与培训、标注审查与核验、标注质量评估等；另一方面，从安全保障来看，生成合成内容的标识涉及个人信息保护、日常合规监管等多种安全需求，重点行业包括数据隐私保护、安全能力评估、安全风险识别、监测和响应等。

1.2《工业互联网安全分类分级管理办法》发布

【内容概述】3 月 20 日工信部发布。《办法》包括总则、企业分类分级、网络安全管理、支持与保障及附则共五章二十二条。《办法》将工业互联网企业分为联网工业企业、平台企业、标识解析企业三类，并依据企业规模、业务范围、应用程度等要素进行自主定级，级别分为三级、二级、一级。《办法》还规定了工业互联网企业的网络安全责任，包括按照相应标准规范落实安全要求、定期开展符合性评测、积极消除网络安全风险等。同时，工信部及地方主管部门需建立健全相关制度机制，加强监测预警、应急处置、安全检查等工作，企业应配合并建立自身网络安全管理制度。

【专家解读】

《工业互联网安全分类分级管理办法》（以下简称《办法》）最早于 2023 年 10 月发布了公开征求意见稿。从本次通知来看，2024 年 4 月就已经正式成文并印发给了地方主管部门和相关企事业单位。

（一）工业互联网和分类分级两大关键词

工业互联网是制造强国和网络强国两大战略的衔接点，而安全则是国家工业互联网专项工作的核心要点之一。近年来，国家相继开展了工业控制系统信息安全防护、工业互联网 + 安全、工业互联网安全深度行等一系列专项行动，持续加强工业互联网安全。

分类分级是重要的网络安全管理思路，当前我国立法明确的分类分级机制包括数据分类分级、工业数据分类分级、工业互联企业网安全分类分级等。而网络安全等级保护制度，则无疑是分类分级机制运作最为成熟的网络安全制度实践之一。

（二）内容及变化

《办法》的核心内容是确定了工业互联网企业的分类及其安全分级标准。同时，还对地方各级工信、通信主管部门监管职责进行了明确。与征求意见稿相比，主要修改体现在两个方面。一是细化了对企业的安全要求，将此前的诸如“建设监测手段”等较为笼统的规定，细化为“监测状态和事件、留存日志、防攻击和病毒”等具体的技术手段要求。二是进一步明确监管边界，增加“涉及关键信息基础设施安全保护的，按照有关规定执行”条款，有助于避免潜在的管理竞合等问题，也有助于减轻对管理对象的重复监管。

（三）思考和影响

总体来看，《办法》是对此前工业互联网企业网络安全分类分级管理试点工作实践的总结与提炼，其重点还是在于强化对工业互联

网企业的安全防护和管理。同时，分级分类只是推进工业互联网安全整体工作的抓手之一，而不是全部，按照《加强工业互联网安全工作的指导意见》等工作规划，后续或将有更多工业互联网安全管理举措推出。如何协调这些不同管理举措之间的合理衔接、避免重复监管，将成为业界关注的焦点问题之一。

从行业影响来看，《办法》的实施对于此前参加过“工业互联网企业网络安全分类分级管理试点工作”的单位和网络安全供应商具有一定利好优势。尤其是对于《工业互联网企业网络安全分类分级管理指南 (试行)》及三类工业互联网企业《网络安全防护规范》的贯标、自主定级、登记备案、检查评估等更加熟悉。此外，工业互联网分类分级工作中的定期符合性评测、安全咨询、安全运维、人员培训等安全服务对网络安全行业也会带来一定的市场增量。

1.3《北京市关于支持信息软件企业加强人工智能应用服务能力行动方案（2025 年）》发布

【内容概述】4 月 8 日北京市经济和信息化局发布。《行动方案》围绕“人工智能 +”战略，推出八大举措支持企业发展。分别是：支持 MaaS 企业在京集聚发展；推动信息软件企业发展行业模型能力；支持通用智能体发展；实施信息软件企业智能技改工程；提升数据治理能力；加速构建开源生态新体系；提升面向中小企业的人工智能服务能力；加强人工智能应用能力培训。

【专家解读】

（一）从背景因素来分析，《行动方案》的发布或主要有两方面的考量

一是促进人工智能的应用推广。2025 年以来，deepseek 的爆火引领了新一轮以国产大模型为主要特征的人工智能应用热潮。国内主要芯片制造商、重点行业用户等都纷纷采取措施加强对国产大模型的支持和部署。在此背景下，如何持续推进和规范引导国产大模型应用的热潮，成为行业尤其是地方主管部门需要思考的重要课题。

二是促进信息软件产业的发展。根据工信部 2024 年软件和信息服务业统计数据，北京市软件产业规模为 3.2 万亿元，占全国（13.7 万亿）的比重超过五分之一，达 22.6%。另据北京市经信局发布的数据，2025 年一季度，信息软件业增加值占全市 GDP 比重达 24.2%，对全市经济增长贡献率接近 4 成，成为北京经济高质量发展“第一引擎”。而推动人工智能大模型的应用，无疑对保持信息软件产业的强劲增长具有重大意义。

（二）从内容来看，《行动方案》主要从三个方面规定了促进应用的主要措施

一是明确了重点鼓励应用的三类人工智能产品和服务形态。《行动方案》第一条至第三条从产品载体形态的角度，分别规定了鼓励应用人工智能全栈式 MaaS 平台、行业大模型、通用智能体三类重点产品和服务形式。对应的鼓励措施分别为“算力券”和部署成本补贴（3000 万元）、“首方案”非硬件部分采购额奖励、算力和模型调用成本支持（3000 万元）。

二是明确了大力发展人工智能所需的两类能力要素。《行动方案》第五条和第六条从鼓励完善人工智能发展要素能力的角度，规定了

大力促进数据治理、开源生态的发展。对应措施分别为“数据券”（50 万元）和共享开源项目资金奖励（200 万元）。

三是明确了促进应用的重要切入点。《行动方案》第四条、第七条、第八条从推进具体工作切入点的角度，规定了信息软件企业“智能技改”、中小企业服务能力、应用能力培训三类具体工作，并明确了智能技改投资奖励（3000 万元）、“中小企业服务券”（20 万元）等鼓励措施。

（三）影响思考

产生的示范性影响。因北京市具有的较为明显的位势和行业占比优势，《行动方案》的发布或将对其他省市产生一定的示范效应。基于支持人工智能大模型国产化趋势，以及促进地方产业发展等重要考量，后续信息软件产业发展排名靠前的相关地方，很有可能也将出台类似的鼓励人工智能应用专项举措。

对行业企业的影响。依托该政策或可争取几个方面的资金或市场支持。一是可结合人工智能大模型的发展成果，积极争取“智能技改”专项资金；二是立足行业特性，打造网络和数据安全“行业模型”典型案例，通过“首方案”渠道争取支持奖励资金等。

1.4《2025 年人工智能技术赋能网络安全应用测试公告》发布

【内容概述】5 月 5 日国家互联网应急中心发布。本次测试活动共设置了 7 个测试场景，包括基于智能体的网络安全自动化分析响应、网络安全告警日志降噪、基于互联网流量的漏洞利用攻击识别及 PoC 生成、基于局域网流量的漏洞利用攻击识别、大模型生成内容安全风险检测、重点车辆船舶监控系统资产脆弱性识别、信用卡异常业务行为检测。

【专家解读】

（一）主要变化

这是该项测试工作开展的第二年，与 2024 年度有几个显著不同。

一是测试的启动方式不同。2025 年度测试通过公开发布公告方式启动，这是与 2024 年度测试相比最大的不同。而 2024 年度的测试工作并未发布公告，主要是通过线下的邀请方式进行。因此，这对于测试工作的知悉度、影响范围、推广效果等方面都可能有较大影响。

二是测试场景的差异。2025 年度设置了 7 个典型场景，与 2024 年度相比，除了“网络安全告警日志降噪”之外，其余 6 个场景均不相同。比较来看，2025 年度测试的通用场景更加侧重于动态网络安全防护，行业场景则在金融基础上增加了交通领域。

三是组织方式更加完善。2025 年度测试工作在组织体系、测试要求和标准、测试流程等方面的规定更加细致、明确，有助于该项测试工作的常态化推进。

此外，公告还对“测试结果应用”进行了说明，从鼓励推广应用的角度明确了测试结果对相关科技奖励、人才评选等的参考价值。

（二）影响思考

近年来，强化网络安全技术产品遴选和推广应用，是有关主管部门一直在大力推进的一项重点工作，试点遴选、典型案例遴选等是此类工作常见的载体方式。与其相比，公开测试的方式更加侧重

技术产品的实际效果，其评价结果的直观性、可量化性、实效性等方面往往具有更强的说服力。

随着该项测试工作规范化、常态化趋势的日益明确，其对于我国网络安全监管体系、行业发展都将产生一定影响。从监管体系看，伴随测试工作相关的协调推进，我国网络安全监管体系在相关工作中的部门协同、职能优化、分工合作等或将更加完善。从行业发展来看，因测试结果具有的潜在应用价值，该项测试无疑将成为人工智能大模型供应商打造品牌影响力的“竞技场”，也将可能成为人工智能大模型用户备货的“购物车”。

2. 国外篇

2.1 欧盟召开第八届网络安全认证大会：庆祝取得的成就并探索未来前景（European Cybersecurity Certification: Celebrating achievements and exploring future horizons）

【内容概述】3 月 13 日欧盟网络安全局发布。会议旨在汇聚网络安全认证生态系统中的利益相关者，共同回顾欧盟网络安全认证重要里程碑，并展望未来的发展和机遇。会上，欧盟网络安全局（ENISA）发布了首批获得欧盟通用标准网络安全认证计划（EU Cybersecurity Certification scheme on Common Criteria，EUCC）认可的合格评定机构，包括 SERMA Safety and Security、Atsec information security GmbH 等共计 10 个，涉及法国、德国、西班牙、瑞典 4 个国家。

【专家解读】

EUCC 是欧盟首个针对信息通信技术（ICT）产品、服务及流程的网络安全认证计划，旨在通过认证提升欧盟范围内 ICT 产品、服务及流程的网络安全水平。该认证计划的法律依据是按照欧盟《网络安全法》（EU 第 2019/881 号）框架制定的《关于采用基于欧盟通用标准的网络安全认证计划的规则》（EU 第 2024/482 号），该规则于 2024 年 1 月 31 日正式生效，并于 2025 年 2 月 27 日开始实施。

欧盟网络安全认证框架下网络安全认证体系主要内容包括认证对象、认证机构、认证级别等 3 个方面。

一是认证对象。主要为进入欧盟的 ICT 产品（网络或信息系统的一个或一组元素）、ICT 服务（通过网络和信息系统传输、存储、检索或处理信息的服务）及 ICT 流程（为设计、开发、交付或维护 ICT 产品或服务而进行的一系列活动）。目前网络安全认证是自愿的，除非欧盟或成员国法律另有规定。

二是认证机构。包括国家网络安全认证机构和符合性评估机构两类。其中，国家网络安全认证机构是由各成员国指定的一个或多个国家网络安全认证机构；符合性评估机构则是根据第 765/2008 号法规（EC）指定的国家专门机构认证认可的第三方评估机构，同时需符合第 2019/881 号法规（EU）附录中规定的合规性要求。

国家网络安全认证机构和符合性评估机构均可开展网络安全认证，并在通过认证后向 ICT 产品、服务和流程的生产商和提供者颁发欧盟网络安全证书。但“高级”保证级别的证书一般只能由国家网络安全认证机构签发，经由国家网络安全认证机构授权的符合性评估机构除外。

三是认证级别。分为基本保证级别 (basic)、充分保证级别 (substantial)、高级保证级别 (high) 三个级别，区别在于认证审查的范围、程序的严格程度。

保证级别	风险评估	审查范围
基本（basic）	对已知事故和网络攻击的可知基本风险进行评估。	技术文件
充分（substantial）	对可知的网络安全风险及因行为者能力有限而导致的事故和网络攻击风险进行评估。	公开的已知漏洞；ICT 产品、服务和流程恰当地实现了必要的安全功能。
高级(high)	对掌握重要技能和资源的行为者造成的高级别网络攻击的风险进行评估。	公开的已知漏洞；证明 ICT 产品、服务和流程在最先进技术背景下实现了必要的安全保障功能；通过“渗透测试”评估其对熟练攻击者的抵抗能力。

按照 ENISA 的相关规划，欧盟网络安全认证未来还将扩展到其他领域，如云服务（EUCS）、5G 网络安全（EU5G）等。

总的来说，EUCC 认证为 ICT 产品进入欧盟市场提供了有力的“通行证”，增强了产品在欧盟市场的竞争力，并且在信息安全领域拥有广泛的认可度。获得 EUCC 认证意味着产品在很大程度上满足了欧盟对于网络安全的高标准要求。

2.2《特朗普总统 2026 财年可自由支配预算请求》（President Trump's Fiscal Year 2026 Discretionary Funding Request）

【内容概述】5 月 2 日美国管理与预算办公室发布。该预算案将非国防可自由支配资金削减 1630 亿美元，比 2025 年制定的水平下降约 23%，这是自 2017 年以来最低的非国防开支水平。同时，该预算案提议国防开支将增加 13%，国土安全部的拨款将增加近 65%，网络安全领域拟削减预算 4.91 亿元。

【专家解读】

该预算案对网络安全领域的预算进行了明显调整，主要体现在对美国网络安全和基础设施安全局（CISA）的预算削减，总金额由 2025 财年的 30 亿美元削减 4.91 亿美元，降至 25.09 亿美元，降幅约为 16%，这是近年来最大幅度的缩减。

特朗普政府 2026 财年网络安全预算调整反映了其“聚焦核心、削减冗余”财政优先级，主要体现在：使 CISA 重新关注其核心职责——联邦网络防御和加强关键基础设施的安全；取消被视为政府“武器化”和浪费性支出的项目；裁撤国际事务办公室；取消与州政府重复的网络安全项目等。与 2025 财年相比，CISA 预算大幅度缩减，而国防和国土安全开支则显著增加，显示出其战略重心向传统安全领域的倾斜。该预算调整或将引发关于国家安全漏洞和效率的争议。

目前，该提案已提交国会审议，最终落地情况可能因两党博弈而有所调整。

NSFOCUS

|巨人背后的专家|

安全加速 国货当"燃"

国产化设备 以旧换新 行动升级

信创自主可控

非信创可更换信创

5大亮点

NF满五年换新

维保加赠

直接让利

THE EXPERT

BEHIND GIANTS

巨人背后的专家

多年以来，绿盟科技致力于安全攻防的研究，
为政府、金融、运营商、能源、交通、科教文卫等行业用户和各类型企业用户，
提供具有核心竞争力的安全产品及解决方案，帮助客户实现业务的安全顺畅运行。
在这些巨人的后面，他们是备受信赖的专家。

客户支持热线：400-818-6868

NSFOCUS绿盟科技