极牛网络安全通讯 • 社群期刊

2025 年 4 月 第 46 期

网安观察



P15
专题：AI大模型部署安全防护实战指南

P16 安全从业者和领导者的 AI风险与安全防护指南
P26 首度揭秘：奇安信AI大模型应用如何防范风险？
P35 如何充分利用通用大模型性能，赋能网络安全

第46期
2025年4月

CONTEN

目录

安全态势

P4 | 《数据安全技术 政务数据处理安全要求》等 6 项网络安全国家标准发布
P4 | 《中华人民共和国卫星导航条例》公开征求意见
P5 | 《网络安全标准实践指南——移动互联网未成年人模式技术要求》发布
P5 | 《关键信息基础设施安全测评要求》公安行业标准发布
P6 | 国家网信办、公安部联合公布《人脸识别技术应用安全管理办法》
P6 | 国家密码管理局《电子认证服务使用密码管理办法》公开征求意见
P7 | 香港通过保护关基设施条例，运营者不更新系统最高可罚 500 万港币
P7 | 工信部印发《工业企业和园区数字化能碳管理中心建设指南》
P8 | 英国政府发布《网络安全与弹性法案》政策声明
P8 | 欧盟发布《数字欧洲计划 2025—2027 年工作规划》
P9 | 哈尔滨市公安局公开通缉 3 名美国国家安全局特工
P9 | 美国工业传感器上市公司森萨塔遭勒索攻击，生产运营被迫中断
P10 | 山东一在校大学生非法获取超 2 万条个人信息，滥用 AI 群发骚扰短信
P10 | 澳大利亚多家大型养老基金超 2 万个账号遭入侵，数百万元养老储蓄金被盗
P11 | 英国皇家邮政 144GB 数据泄露，供应商 Spectos 疑似“罪魁祸首”
P11 | OA 系统遭境外入侵窃取信息，四川南充一行政单位被处罚
P12 | OA 系统漏洞致使数据泄露，青海某公司被罚 5 万元
P13 | Foxmail for Windows 远程代码执行漏洞安全风险通告
P14 | Vite 任意文件读取漏洞 (CVE-2025-31125) 安全风险通告
P14 | CrushFTP 身份验证绕过漏洞安全风险通告



国际视野

P8
英国 NCSC 发布《迁移到后量子密码学的时间表》指南

ONTENTS



专题报道

P16
安全从业者和领导者的 AI风险与安全防护指南

P26
首度揭秘：奇安信AI大模型应用如何防范风险？

P35
如何充分利用通用大模型性能，赋能网络安全



第46期

《网安观察》编辑部

主办　极牛网

总　　编　　辑：陈鑫杰
总　　顾　　问：叶绍琛
副　　总　　编：王文彦
威胁情报主编：陈艇鑫
移动安全主编：蔡国兆
网安人才主编：林俊濠
涉网犯罪主编：胡铭凯
网安产业主编：张九史
网安态势主编：郑泽彬

公众号　小程序　网站

电子版请访问 www.geeknb.com 阅读或下载索阅、投稿、建议和意见反馈，请联系极牛网期刊编辑部。

E m a i l：hi@geeknb.com
地　　址：深圳市龙岗区天安云谷2栋2层
邮　　编：518000
电　　话：0755-33228862
印刷数量：1000 本
印刷单位：深圳彩虹印刷有限公司

版权所有 ©2021 极牛网，保留一切权利。
非经极牛网书面同意，任何单位和个人不得擅自摘抄、复制本资料内容的部分或全部，并不得以任何形式传播。

无担保声明
本资料内容仅供参考，均“如是”提供，除非适用法要求，极牛网对本资料所有内容不提供任何明示或暗示的保证，包括但不限于适销性或者适用于某一特定目的的保证。在法律允许的范围内，极牛网在任何情况下都不对因使用本资料任何内容而产生的任何特殊的、附带的、间接的、继发性的损害进行赔偿，也不对任何利润、数据、商誉或预期节约的损失进行赔偿。

政策篇



国内，各行业数字化安全不断完善。《卫星导航条例》公开征求意见，要求加强卫星导航运行安全、网络安全、应用安全、数据安全保护；国家能源局《2025年电力安全监管重点任务》提出，评估调整国家级电力网络安全靶场，推动行业漏洞库等基础设施建设；中央两办印发《关于健全社会信用体系的意见》，要求建立信用信息安全管理追溯和侵权责任追究机制。

国际上，欧洲各国深化落实欧盟网络安全法律要求。欧盟发布《数字欧洲计划2025—2027年工作规划》，将投入资金开发《网络弹性法案》单一报告平台；瑞典《国家网络安全战略2025—2029年》提出，推动欧盟《网络弹性法》的实施；英国政府拟推出《网络安全与弹性法案》，对接NIS2指令要求加强关基保护。

国内

《数据安全技术 政务数据处理安全要求》等6项网络安全国家标准发布

4月9日，根据2025年3月28日国家市场监督管理总局、国家标准化管理委员会发布的中华人民共和国国家标准公告（2025年第6号），全国网络安全标准化技术委员会归口的6项国家标准正式发布。具体包括《网络安全技术 运维安全管理产品技术规范》《数据安全技术 政务数据处理安全要求》《数据安全技术 基于个人信息的自动化决策安全要求》《数据安全技术 数据安全评估机构能力要求》《数据安全技术 大型互联网企业内设个人信息保护监督机构要求》《网络关键设备安全技术要求 可编程逻辑控制器（PLC）》。

《中华人民共和国卫星导航条例》公开征求意见

4月3日，国家发展改革委牵头会同有关部门研究起草了《中华人民共和国卫星导航条例（公开征求意见稿）》，现向社会公开征求意见。该文件共7章52条，分为总则、建设运行、运营服务、应用推广、安全保护、法律责任和附则。该文件提出，国家建立健全卫星导航安全体系，加强卫星导航运行安全、网络安全、应用安全、数据安全保护，确保卫星导航安全。国家有关部门负责监测、防御、处置来源于境外的风险和威胁，保护相关系统免受攻击、侵入、干扰和破坏；国家网信部门负责卫星导航应用的网络安全和数据安全统筹协调；国务院有关部门和县级以上地方人民政府，应当按照职责分工，负责本部门、本行政区域卫星导航应用的网络安全和数据安全管理工作，督促落实网络安全和数据安全有关制度。

《工业和信息化部关于开展号码保护服务业务试点的通知》公开征求意见

4月3日，工业和信息化部信息通信管理局编制了《工业和信息化部关于开展号码保护服务业务试点的通知（征求意见稿）》，现公开征求意见。该文件指出，号码保护服务是互联网平台等企业为保护用户的个人电话号码信息而提供的一项服务，也称“中间号”或“隐私号”。该文件明确了业务定义和业务参与方，明晰了应用平台提供方、基础电信业务经营者、业务使用方等三方参与主体的责任边界和相关要求，强化合规经营、防范治理电信网络诈骗、加强商业营销电话和商业营销信息防控等业务管理要求，限制应用平台提供方“转租转售”。该文件规划了15位长的700专用号码，并开展为期两年半的试点行动，试点过渡阶段结束后，将全部使用700专用号码开展号码保护服务业务。

《网络安全标准实践指南——移动互联网未成年人模式技术要求》发布

4 月 3 日，全国网络安全标准化技术委员会秘书处组织编制了《网络安全标准实践指南——移动互联网未成年人模式技术要求》。该文件规定了移动互联网未成年人模式的技术要求，适用于移动互联网应用程序提供者、移动智能终端制造商和移动互联网应用程序分发平台提供者开展未成年人模式的研发和应用，也可为监管部门、第三方评估机构对未成年人网络保护的监督、管理、评估提供参考。

《关键信息基础设施安全测评要求》公安行业标准发布

3 月 31 日，公安部技术监督委员会 2024 年 12 月 26 日批准发布了公安部网络安全保卫局组织起草的《关键信息基础设施安全测评要求》（GA/T2182-2024）行业标准，自 2025 年 5 月 1 日起实施。该标准规定了针对关键信息基础设施开展安全测评的要求和方法，适用于规范关基运营者及网络安全服务机构开展关基安全测评工作。关基测评的目标是发现关基现存的或潜在的安全风险，为运营者开展建设整改、提升其关基安全保护水平及安全管控能力提供指导。关基测评由单元测评、关联测评和整体评估三部分组成。关基测评在开展整体评估时复用相关等级测评结果。该标准是推动关基安全保护、建立全面安全防护体系的重要手段，是各级公安机关推进关基安全保护工作的重要依托。

中共中央办公厅、国务院办公厅印发《关于健全社会信用体系的意见》

3 月 31 日，中共中央办公厅、国务院办公厅印发《关于健全社会信用体系的意见》。该文件专设单章加强信用信息安全保护，要求建立信用信息安全管理追溯和侵权责任追究机制，明确信息传输链条各环节安全责任。严格落实安全保护责任，规范信用信息处理程序。提高信用信息基础设施安全管理水平，建立健全信用信息安全应急处理机制。

《中华人民共和国网络安全法（修正草案再次征求意见稿）》公开征求意见

3 月 28 日，国家互联网信息办公室会同相关部门进一步研究起草了《中华人民共和国网络安全法（修正草案再次征求意见稿）》，现向社会公开征求意见。该文件修改思路主要包括：重点强化网络安全法律责任并加大对违法行为处罚力度；加强与《数据安全法》《个人信息保护法》《行政处罚法》等相关法律有机衔接；科学设置网络运行安全、网络信息安全等不同类型违法行为的法律责任。该文件还提出，网络运营者存在主动消除或者减轻违法行为危害后果、违法行为轻微并及时改正，且没有造成危害后果或者初次违法且危害后果轻微并及时改正等情形的，依法从轻、减轻或者不予处罚。

四部委发布《关于开展 2025 年个人信息保护系列专项行动的公告》

3 月 28 日，中央网信办、工业和信息化部、公安部、市场监管总局发布《关于开展 2025 年个人信息保护系列专项行动的公告》，进一步深入治理常用服务产品和常见生活场景中存在的违法违规收集使用个人信息典型问题，切实维护人民群众在网络空间合法权益。专项行动将主要围绕六方面展开，包括 App（含小程序、公众号、快应用）、SDK、智能终端、线下消费场景违法违规收集使用个人信息，公共场所违法违规收集使用人脸识别信息，个人信息相关违法犯罪案件。专项行动将重点打击通过暗网电报等境外渠道及境内渠道，违规售卖公民个人信息等违法犯罪案件。

《工业和信息化领域人工智能安全治理标准体系建设指南（2025）（征求意见稿）》公开征求意见

3 月 27 日，工业和信息化部人工智能标准化技术委员会秘书处编制了《工业和信息化领域人工智能安全治理标准体系建设指南（2025）（征求意见稿）》，现面向社会公开征求意见。该文件提出，人工智能安全治理标准体系结构包括治理能力、基础设施安全、网络安全、数据安全、算法模

型安全、应用安全、赋能安全等 7 个部分。该文件指出，截至 2025 年 2 月，在人工智能安全治理领域，国内共发布了 55 项国家标准与行业标准，在未来三年内，针对 AI 治理能力、AI 数据安全、AI 算法模型、AI 应用安全等多个方面，还计划发布 70 个国家标准和行业标准，进一步完善 AI 安全治理标准体系。



国家网信办、公安部联合公布《人脸识别技术应用安全管理办法》

3 月 21 日，国家互联网信息办公室、公安部联合公布《人脸识别技术应用安全管理办法》，自 2025 年 6 月 1 日起施行。该文件共 20 条，对应用人脸识别技术处理人脸信息的基本要求和处理规则、人脸识别技术应用安全规范、监督管理职责等作出了规定。该文件要求，人脸识别技术应用系统应当采取数据加密、安全审计、访问控制、授权管理、入侵检测和防御等措施，保护人脸信息安全。涉及网络安全等级保护、关键信息基础设施的，应当按照国家有关规定履行网络安全等级保护、关键信息基础设施保护义务。个人信息处理者应当在应用人脸识别技术处理的人脸信息存储数量达到 10 万人之日起 30 个工作日内，向所在地省级以上网信部门履行备案手续。



国家密码管理局《电子认证服务使用密码管理办法》公开征求意见

3 月 21 日，国家密码管理局研究起草了《电子认证服务使用密码管理办法（征求意见稿）》，现公开征求意见。该文件为《电子认证服务密码管理办法》的修订稿，包括总体要求、许可审批、运行规范、监督检查、法律责任和附则 6 个方面的内容，共 23 条。该文件要求，电子认证服务机构应当建立健全电子认证服务系统运行维护制度，明确关键岗位和岗位责任，制定操作规范，加强巡检和运行维护，提高监测预警和应急处置能力。电子认证服务机构应当自行或委托专业机构每年至少进行一次电子认证服务使用密码合规性评估，每年组织开展电子认证服务使用密码安全知识和岗位操作技能培训，相关技术人员和管理人员每年教育培训时间不得少于 20 个小时。



国家能源局印发《2025 年电力安全监管重点任务》

3 月 21 日，国家能源局综合司印发《2025 年电力安全监管重点任务》。该文件提出了 4 个方面、共 22 项重点任务，其中包括完善网络安全风险管控体系。该文件要求，研究修订电力监控系统安全防护总体方案等安全防护方案和评估规范，完善电力监控系统安全防护政策体系。组织开展电力行业网络安全演习，评估调整国家级电力网络安全靶场，推动行业漏洞库等基础设施建设。动态开展电力行业关键信息基础设施认定，推进电力行业网络安全等级保护定级审核，编制“十五五”能源关键信息基础设施安全规划。



工信部公布《工业互联网安全分类分级管理办法》

3 月 20 日，工业和信息化部官网公布《工业互联网安全分类分级管理办法》全文。该文件于 2024 年 4 月 11 日印发。该文件共 5 章 22 条，包括总则、企业分类分级、网络安全管理、支持与保障、附则。该文件提出，工业互联网安全分类分级管理工作遵循统筹指导、分类施策、分级防护、突出重点的原则，指导企业提升安全防护能力。该文件要求，工业互联网企业应当按照工业互联网安全定级相关标准规范，结合自身情况开展自主定级。工业互联网企业级别由高到低分为三级、二级、一级，三级工业互联网企业每年至少开展一次评测，二级工业互联网企业每两年至少开展一次评测，一级工业互联网企业可参照二级企业相关要求开展评测。



《信息安全管理体系审核和认证机构要求》等 3 项国家标准公开征求意见

3 月 20 日，全国网络安全标准化技术委员会归口的 3 项国家标准现已形成标准征求意见稿，现公开征求意见。3 项标准包括：《网络安全技术 信息安全管理体系审核和认证机构要求》在 GB/T 27021.1—2017 的基础上，对信息安全管理体系审核和认证机构规定了要求并提供了指南；《网络安全技术 网络安全事件管理 第 1 部分 原理和过程》提出了网络安全事件管理关键活动的基本概念、原理和过程，提供了结构化的方法来准备、发现、报告、评估和响应事件并

进行经验总结；《网络安全技术　网络安全事件管理　第2部分：事件响应规划和准备指南》基于GB/T 20985.1的6.2和6.6中给出的“网络安全事件管理阶段”模型的“规划和准备”阶段和“经验总结”阶段，给出了事件响应规划和准备及事件响应经验总结的指南。

香港通过保护关基设施条例，运营者不更新系统最高可罚500万港币

3月19日，香港立法会正式通过《保护关键基础设施（电脑系统）条例草案》。该法案涵盖被认为对社会正常运作至关重要的八大领域，包括能源、信息技术、银行及金融服务、海运、陆路运输、航空运输、医疗服务和通信及广播服务。此外，其他涉及关键社会及经济活动的基础设施运营者，如管理大型体育或表演场馆、科研园区的机构也被纳入法案范围。该法案要求，关键基础设施运营者若未能确保关键计算机系统的安全更新，最高将被罚款500万港币。香港特区政府保安局局长邓炳强表示，当局计划于今年6月成立专员办公室，并筛选受影响的运营者，目标是在2026年1月1日前使法案正式生效。

工信部印发《工业企业和园区数字化能碳管理中心建设指南》

3月18日，工业和信息化部印发《工业企业和园区数字化能碳管理中心建设指南》。该文件包括建设目标、业务功能、技术方案、组织保障四部分。该文件提出，工业企业和园区应增强网络和数据安全保护意识，落实《信息安全技术 网络安全等级保护基本要求》（GB/T 22239）等国家标准，压实网络和数据安全主体责任。根据实际情况，对能碳管理中心设定相应的安全等级保护级别，做好重要数据识别、分级防护和风险评估，保障数据安全。

《网络攻击和网络攻击事件判定准则》等2项网络安全国家标准发布

3月17日，根据2025年2月28日国家市场监督管理总局、国家标准化管理委员会发布的中华人民共和国国家标准公告（2025年第4号），全国网络安全标准化技术委员会归口的2项国家标准正式发布。具体清单包括《网络安全技术 公钥基础设施 在线证书状态协议》《网络安全技术 网络攻击和网络攻击事件判定准则》。

工信部印发《关于做好2025年信息通信业安全生产和网络运行安全工作的通知》

3月17日，工业和信息化部印发《关于做好2025年信息通信业安全生产和网络运行安全工作的通知》。该文件包括总体要求、主要任务、保障措施三部分，要求按照“三管三必须”要求，进一步树牢安全发展理念，着力提升行业本质安全水平，坚决防范遏制重特大事故发生。该文件明确了加强思想政治引领、抓好通信建设安全生产制度落实、强化信息通信网络运行安全管理、进一步压实企业安全生产主体责任、加强事故隐患排查治理、突出重点领域安全管理、提高应急处置能力等七项主要任务。

国家档案局办公室印发《推进数字档案馆建设实施办法（试行）》

3月14日，国家档案局办公室印发《推进数字档案馆建设实施办法（试行）》。该文件提出，数字档案馆建设主要包括基础设施、应用系统、档案数字资源、档案共享利用、制度规范、安全保障等内容。该文件要求，近3年发生过重大信息安全事故或存在严重信息安全隐患的，不能申请数字档案馆认定。国家档案局和省、自治区、直辖市档案主管部门应当对各自认定的数字档案馆进行监督检查，及时向存在工作停滞、安全隐患突出等严重问题的数字档案馆提出整改要求，未按时、按要求完成整改的可撤销认定结果。

国际

欧盟EDPB发布《人工智能隐私风险和缓解指南：大语言模型》

4月10日，欧洲数据保护委员会（EDPB）发布了由

第三方专家撰写的《人工智能隐私风险和缓解指南：大语言模型》报告。该文件提出了一套全面的大语言模型系统风险管理方法，以系统地识别、评估和缓解隐私与数据保护风险，并针对大语言模型系统的常见隐私风险，提出了一系列可实践的缓解措施。该文件还提供了三个具体场景的应用案例，包括用于客户询问的聊天机器人、监控和支持学生改进的大语言模型系统、旅行和日程管理 AI 助手。



英国政府发布《网络安全与弹性法案》政策声明

4 月 1 日，英国政府发布网络安全与弹性政策声明，详细介绍了将于今年提交议会的《网络安全与弹性法案》内容。该政策声明显示，《网络安全与弹性法案》主要内容包括：对接欧盟 NIS2 指令要求，扩大关基行业覆盖范围，提高关基网络安全事件上报标准；将约一千家托管服务商纳入管控范围，要求遵循必要的网络安全标准，以提高 IT 基础设施的安全性；加强供应链安全，可对基础服务运营商、相关数字服务提供商及监管机构单独给出的指定关键供应商，设置更严格的供应链安全责任。



欧盟发布《数字欧洲计划 2025—2027 年工作规划》

3 月 28 日，欧盟委员会发布了《数字欧洲计划 2025—2027 年工作规划》。该文件重点关注人工智能、云计算、数据、网络安全等领域，支持欧盟成为“人工智能大陆”的战略目标。该文件将投入 4560 万欧元用于网络安全领域，包括建立欧盟网络安全储备、开发《网络弹性法案》单一报告平台及发展网络安全技能学院等。该文件还将投入 3.672 亿欧元用于欧洲网络安全能力中心，重点投资人工智能、后量子转型等网络安全新技术。该文件反映了欧盟在技术主权、民主和安全方面的政治优先事项，将为欧洲数字化转型提供强有力的支持，推动欧洲在全球数字创新舞台上的领先地位。



英国信息专员办公室发布匿名化指南

3 月 28 日，英国信息专员办公室（ICO）发布了《匿名化指南》，旨在帮助组织有效实施匿名化技术，确保个人数据在共享、发布或再利用时，符合英国 GDPR、2018 年数据保护法等数据保护法规。该文件指出，将个人数据应用匿名化技术转化为匿名信息的行为，算作处理个人数据。虽然最终结果（匿名信息）不受数据保护法约束，但该过程（匿名化）是受约束的。该文件提出了“有动机的闯入者”测试作为评估匿名化有效性的方法，即模拟具备中等技术能力、资源支持且存在明确动机（如经济利益、社会关注或学术研究）的攻击者，通过综合运用公开数据源（如社交媒体、政府开放数据）、技术手段（如数据关联分析、算法攻击）及社会工程学方法，尝试破解匿名化数据集，以重新识别个体身份。



瑞典政府发布《国家网络安全战略 2025—2029 年》

3 月 20 日，瑞典政府正式向议会提交了《国家网络安全战略 2025—2029 年》，对未来五年的国家网络安全建设进行了整体谋划，以应对日益复杂的网络安全威胁，提升国家整体网络安全水平。该文件提出了三大支柱 13 项具体目标，包括系统高效的网络安全工作、提升网络安全知识和能力、加强网络安全事件的预防处置。该文件还明确了当前面对的重要挑战及应对举措，如针对敌对国家威胁，要求进一步增加关键基础设施的网络安全防护水平和主动防御能力；针对网络犯罪要求重点打击勒索软件和数据盗窃，增强执法部门的技术工具；供应链方面，要求相关企业评估供应商风险，推动欧盟《网络弹性法》的实施。



英国 NCSC 发布《迁移到后量子密码学的时间表》指南

3 月 20 日，英国国家网络安全中心（NCSC）发布《迁移到后量子密码学的时间表》指南，为各组织在接下来数年安全迁移到后量子密码学提供了清晰的路线图和关键时间节点。该文件提出了三大阶段为期 10 年的迁移计划，到 2028 年实现发现与评估、初步计划制定，到 2031 年实现高优先级迁移活动执行，到 2035 年实现全面迁移完成，并针对每个阶段给出了详细操作指南。



事件篇

我国多部门开展网络攻击溯源调查活动。国家计算机病毒应急处理中心发布报告称，哈尔滨亚冬会赛事信息系统遭境外网络攻击超 27 万次，攻击源大部分来自美国，随后哈尔滨市公安局宣布成功追查到美国国家安全局的 3 名特工和两所美国高校参与实施了攻击活动，并公开通缉 3 名特工；国家安全部发文披露台湾资通电军充当“台独”分裂势力的爪牙，持续对大陆开展网络攻击渗透活动，并直接公布了 4 名“台独”网军身份信息。

哈尔滨市公安局公开通缉 3 名美国国家安全局特工

4 月 15 日新华社消息，黑龙江省哈尔滨市公安局宣布，为依法严厉打击境外势力对我网络攻击窃密犯罪，切实维护国家网络空间安全和人民生命财产安全，决定对 3 名隶属于美国国家安全局（NSA）的犯罪嫌疑人凯瑟琳·威尔逊（Katheryn A. Wilson）、罗伯特·思内尔（Robert J. Snelling）、斯蒂芬·约翰逊（Stephen W. Johnson）进行通缉。前期，“2025 年哈尔滨第九届亚冬会”遭受境外网络攻击事件经媒体报道后，引发广泛关注。国家计算机病毒应急处理中心和亚冬会赛事网络安全保障团队，及时向哈尔滨市公安局提交了亚冬会遭受网络攻击的全部数据。哈尔滨市公安局立即组织技术专家组成技术团队开展网络攻击溯源调查。在相关国家支持下，经技术团队持续攻坚，成功追查到 NSA 的 3 名特工和两所美国高校，参与实施了针对亚冬会的网络攻击活动。调查发现，NSA 意图利用网络攻击窃取参赛运动员的个人隐私数据，并妄图破坏系统，扰乱影响亚冬会赛事的正常运行。同时，NSA 针对黑龙江省内能源、交通、水利、通信、国防科研院校等重要行业开展网络攻击，意图破坏我关键信息基础设施，引发社会秩序混乱和窃取我相关领域重要机密信息。进一步调查发现，该 3 名特工曾多次对我国关键信息基础设施实施网络攻击，并参与对华为公司等企业的网络攻击活动。技术团队同时发现，具有 NSA 背景的美国加利福尼亚大学、弗吉尼亚理工大学也参与了本次网络攻击。

美国工业传感器上市公司森萨塔遭勒索攻击，生产运营被迫中断

4 月 10 日 Bleeping Computer 消息，美国上市公司、知名传感器生产商森萨塔科技（Sensata Technologies）遭遇勒索软件攻击，公司网络内的部分设备遭到加密，致使业务运营中断。森萨塔科技在 SEC 披露文件中表示，此次攻击发生于 4 月 6 日（星期日），涉及数据被窃取的情况。此次事件已对森萨塔科技的运营造成暂时影响，包括发货、收货、制造生产及其他多个支持职能。森萨塔科技表示，公司已立即采取措施，加快受此次网络攻击影响的关键功能的恢复工作，不过目前尚无法提供完整的恢复时间表。公司表示，预计本季度的财务业绩不会受到实质性影响。

摩洛哥国家社保基金遭网络攻击，近 200 万民众敏感数据或泄露

4 月 9 日 Yabiladi 消息，摩洛哥国家社保基金疑似发生重大数据泄露事件，超过 5.4 万份文件被盗，近 200 万企业员工的敏感数据遭泄露，其中涉及姓名、身份证号码、公司关系、邮件地址、电话号码、银行账户、工资申报记录等。该机构声明称，初步调查显示，此次泄密是由于攻击者绕过了安全系统。声明未透露攻击者的身份，但表示对方发布的许多文件具有误导性、不准确性或不完整性。此前，疑似阿尔及利亚黑客组织 JabaRoot DZ 在 Telegram 和暗网泄露论坛上发布了相关数据。

山东一在校大学生非法获取超 2 万条个人信息，滥用 AI 群发骚扰短信

4 月 7 日公安部网安局消息，山东公安网安部门侦破一起非法获取计算机信息系统数据案，犯罪嫌疑人非法获取两万余条学生个人信息，后利用 AI 技术向其中的两千余名学生发送骚扰短信。犯罪嫌疑人胡某是一名在校大学生，非法入侵了学校某系统并获取了两万余条该校学生个人信息。为寻求刺激、炫耀技术，胡某通过之前发现的某小程序存在的技术漏洞，利用 AI 编写程序，把其中盗取的上千余名学生的手机号码在该小程序上批量注册账户，后将短信验证码篡改为淫秽内容发送至学生本人，对其进行短信骚扰。目前，犯罪嫌疑人对非法获取计算机信息系统数据罪供认不讳，案件正在进一步侦办中。

澳大利亚多家大型养老基金超 2 万个账号遭入侵，数百万元养老储蓄金被盗

4 月 5 日路透社消息，据知情人士透露，黑客对澳大利亚主要的养老基金发起了一系列协同攻击。在 3 月 29 日至 30 日（周末）期间，多家大型养老基金超过 2 万个用户账号遭到入侵，其中仅 AustralianSuper 基金用户就有至少 50 万澳元（约合人民币 220 万元）的资金被转出。目前整体损失规模尚不确定，AustralianSuper、Australian Retirement Trust、Rest、Insignia 及 Hostplus 等知名基金均确认受到影响。澳大利亚国家网络安全协调员 Michelle McGuinness 发布声明称，网络犯罪分子正以澳大利亚 4.2 万亿澳元退休储蓄领域的账号为目标，目前正协调政府、监管机构及相关行业作出应对。

美国巴尔的摩市政府遭社工攻击，超 1100 万元供应商合同款被盗

4 月 3 日 StateScoop 消息，美国马里兰州首府巴尔的摩市审计长办公室透露，3 月份发生的一起网络攻击事件导致该市遭遇身份盗窃欺诈，损失超过 150 万美元（约合人民币 1100 万元）。副审计长 Erika McClammy 表示，3 月 13 日，该市收到通知称，一名网络犯罪分子攻击了市政府的应付账款部门，并通过身份盗窃手段获取了超过 150 万美元的资金，这笔款项原计划支付给一名市政供应商。McClammy 表示：“我们目前尚未确定幕后黑手的身份。显然，对方很可能使用了多个身份。他们成功绕过了巴尔的摩市的地理围栏系统，使用了（星链）IP 地址。”官方报告指出，犯罪分子从去年秋季开始与市政府建立联系，利用网络上公开的信息，冒充一名现有供应商的员工，与多个市政部门建立了联系，从而逐步渗透进市政系统。

哈尔滨亚冬会赛事信息系统遭境外网络攻击超 27 万次，攻击源大部来自美国

4 月 3 日央视新闻消息，国家计算机病毒应急处理中心发布“2025 年哈尔滨第九届亚冬会”赛事信息系统及黑龙江省内关键信息基础设施遭境外网络攻击情况监测分析报告，全面总结了网络安全保障团队监测处置的本届赛事各类网络安全威胁情况。相关统计数据表明，赛事期间，各赛事信息系统、黑龙江省域范围内的关键信息基础设施遭到来自境外的大量网络攻击。攻击源大部分来自美国、荷兰等国家和地区。在赛事网络安全保障团队的共同努力下，这些网络攻击未能对赛事造成严重影响，但却进一步凸显了我国网络频繁遭受境外攻击的严峻形势。

俄乌黑客展开铁路大战，运营系统瘫痪殃及百万民众

4 月 1 日综合消息，俄乌数字战争近期在交通行业升级，导致数以百万计的民众出行受阻。3 月下旬，乌克兰国有铁路运营商 Ukrzaliznytsia 遭遇了一场“系统性、复杂且多层次”的网络攻击。这次攻击直接导致在线售票系统宕机，网站和移动应用瘫痪。虽然火车运行时间表未受影响，但乘客不得不涌向车站窗口排队购票，场面一度混乱。为了应对危机，Ukrzaliznytsia 紧急加倍增开售票窗口，并在 89 小时的不懈努力后恢复了在线服务。3 月 31 日，莫斯科地铁系统的网站和应用突然宕机。根据俄罗斯用户在 Downdetector 上的反馈，这次故障不仅让在线服务中断，连数字交通卡都无法正常使用，数百万民众出行受到影响。更戏剧性的是，在莫斯科地铁网站被重新定向到了乌克兰国有铁路运营商 Ukrzaliznvtsia 的网站，页面上赫然出现了一条 Ukrzaliznytsia 网站被攻击后黑客留下的消息：“你们打我们的铁路，我们就瘫痪你们的地铁。”

英国皇家邮政 144GB 数据泄露，供应商 Spectos 疑似“罪魁祸首”

4 月 2 日 Hackread 消息，英国皇家邮政集团日前遭遇了一起严重的数据泄露事件，144GB 的敏感数据被黑客公开，包括客户个人信息、内部通讯记录、运营数据及营销基础设施数据等。一位名为 GHNA 的黑客在地下论坛 Breach Forum 上发布了相关数据，并声称是通过皇家邮政的供应商 Spectos 获取的这些信息。泄露的数据包包含 293 个文件夹和 16,549 个文件，涉及客户个人身份信息、内部通讯记录、运营数据、营销基础设施数据等，Spectos 的名字多次出现在泄露材料中，包括内部文件和录制的视频通话。皇家邮政已承认此事，并表示正在与 Spectos 合作调查。

OA 系统遭境外入侵窃取信息，四川南充一行政单位被处罚

4 月 1 日，南充市互联网信息办公室近日依法查处一起违反《中华人民共和国网络安全法》典型案例。经查，该市某行政单位 OA 系统存在严重安全隐患：未按要求完成网络安全等级保护备案；系统日志留存周期不足法定六个月标准；系统存在“弱口令”漏洞。上述问题直接导致境外黑客组织成功入侵系统窃取信息。该单位涉嫌违反《中华人民共和国网络安全法》第二十一条第（三）项、第（四）项规定。市互联网信息办公室依据《中华人民共和国网络安全法》第五十九条第一款规定，依法对该单位作出警告并责令改正的行政处罚。

外包保洁员主动投靠境外情报机关，造成重大失泄密

3 月 28 日国家安全部公众号消息，国家安全部发文称，服务外包给涉密机关、单位带来便利的同时，也暗藏着失泄密隐患。某重要涉密单位长期将办公楼物业管理外包给某物业服务公司运营，并未对该公司开展必要的保密监管。该公司保洁人员段某为满足个人私欲，主动联系境外间谍情报机关，后经指使借从事保洁服务之机，通过盗取、偷拍等方式为境外间谍情报机关提供 1 项机密级、2 项秘密级国家秘密及其他 6 项情报，造成重大失泄密，危害我国家安全。

疑似 NSA 网攻行动曝光：神秘 0day 漏洞利用链 目标针对俄媒体科研机构

3 月 27 日 The Record 消息，卡巴斯基披露了一起尖端网络攻击行动，受害者点击定向钓鱼邮件中的链接，该页面暗藏了零 0day 漏洞（CVE-2025-2783）利用代码，可远程绕过 Chrome 浏览器的沙盒保护机制，结合另一个未知漏洞即可实现远程代码执行，控制受害者的设备。根据该行动非常隐蔽和技艺高超的特征，卡巴斯基认为攻击者具有国家背景，参考此前披露的三角行动，这起事件也很可能是 NSA 的网络间谍活动。

600 万 Oracle 云用户数据疑泄露，官方否认

3 月 26 日 Bleeping Computer 消息，一位自称“rose87168”的黑客宣称成功入侵了 Oracle Cloud 的联邦单点登录服务器，窃取了 600 万用户的认证数据和加密密码，并开始在暗网兜售这些信息。Oracle 官方矢口否认遭遇任何入侵。不过外媒 BleepingComputer 通过多方验证，认为黑客提供的数据样本属实，多家公司在匿名承诺下确认，包括 LDAP 显示名称、邮箱地址、姓名等身份信息在内，这些数据完全属实，与其内部记录一致。

网络攻击严重扰乱运营，南非家禽养殖上市公司预告利润大幅下滑

3 月 24 日 TheCyberExpress 消息，南非上市家禽生产商、最大的鸡肉生产商 Astral Foods 发布披露报告称，一起网络安全事件严重扰乱了公司运营，预计将导致截至 2025 年 3 月 31 日的半个财年内，产生约 2000 万兰特（约合人民币 795 万元）的利润损失。Astral Foods 表示，该事件发生在 3 月 16 日，导致公司家禽加工部门暂时停工。此次停工延误了加工和交付，直接影响了收入。尽管公司迅速启动了灾难恢复预案，但暂时的停工仍导致了财务损失。南非家禽养殖业正面临多重挑战，受该事件与其他因素影响，

Astral Foods 预计半年报净利润最多将下滑 60%。该公司表示采取了措施降低影响，已恢复正常运营。

OA 系统漏洞致使数据泄露，青海某公司被罚 5 万元

3 月 21 日网信青海公众号消息，青海省互联网信息办公室依法对海西州某公司存在的网络数据安全违法行为进行行政处罚，开出该省首张网络数据安全“罚单”。青海省互联网信息办公室针对有关问题线索充分调查取证，依法查明海西州某公司存在不履行网络安全、数据安全保护义务行为，其办公 OA 系统未采取技术措施和其他必要措施保障数据安全，存在未授权访问漏洞，造成企业部分数据泄露。省互联网信息办公室依据《中华人民共和国数据安全法》第四十五条第一款规定，对该公司作出责令改正，给予警告，并处以 5 万元罚款的行政处罚，对直接主管人员和其他责任人员分别处以 1 万元罚款处罚。

美国最大精子库数据泄漏，或引发全球生殖医疗隐私危机

3 月 18 日 Bleeping Computer 消息，美国最大精子库加州冷冻银行（California Cryobank）向客户发出警告称，其系统发生数据泄露事件，客户姓名、社保号、银行账户等敏感信息或遭窃取，可能对全球数十万通过人工生殖技术诞生的后代隐私造成长期威胁。该公司披露称，黑客在 2024 年 4 月 20 日至 22 日期间入侵系统，窃取了包含客户姓名、社保号、驾照号、银行账户及路由码、信用卡信息、健康保险资料等在内的多维度隐私数据。该公司表示，向社保号或驾照号泄露的用户提供了免费一年的信用监控服务。

韩国区块链游戏平台 WEMIX 遭黑客攻击，损失 622 万美元

3 月 18 日 Bleeping Computer 消息，韩国区块链游戏公司 WEMIX 披露，公司于 2 月 28 日遭受黑客攻击，导致约 865 万枚 WEMIX 代币被盗，价值约 622 万美元。黑客通过窃取 NFT 平台“NILE”的认证密钥，成功进行了多次提取操作，其中 13 次成功，将被盗代币转移至多个钱包，并迅速在海外交易所变现。此次黑客攻击事件对 WEMIX 代币价格造成了严重影响，导致其价格从 940 韩元骤跌至 620 韩元区间，跌幅超过 30%。为弥补投资者损失，WEMIX 基金会于 3 月 13 日宣布了 100 亿韩元回购计划，并于次日追加 2000 万枚 WEMIX 代币回购。同时，WEMIX 正在努力恢复服务，计划于 3 月 21 日全面恢复，目前所有区块链相关基础设施正在迁移到更安全的环境。

Fortinet 防火墙漏洞遭勒索软件利用，全球多家企业被黑

3 月 17 日 TechCrunch 消息，安全厂商 Forescout Research 发布报告称，与臭名昭著的 LockBit 团伙有关的黑客正利用 Fortinet 防火墙的两个漏洞，在多家企业网络中部署勒索软件。该公司研究人员表示，自 2024 年 12 月以来，攻击者先后利用了 CVE-2024-55591、CVE-2025-24472 等漏洞，对 Fortinet 客户的企业网络发起攻击。Fortinet 已于今年 1 月发布了针对这两个漏洞的安全补丁。Forescout 威胁狩猎高级经理 Sai Molige 表示，该公司已调查了三起不同企业遭遇的入侵事件，但他们认为实际受害者可能远不止这些。在其中一起已确认的攻击事件中，Forescout 发现，攻击者会“有针对性地”加密存储敏感数据的文件服务器。

国家安全部直接公布 4 名“台独”网军身份信息

3 月 17 日国家安全部消息，国家安全部公众号发文称，台湾资通电军自 2017 年 6 月成立以来，便充当“台独”分裂势力的爪牙，无所不用其极地对大陆开展网络攻击渗透活动。国家安全机关严密监测，深入调查，已锁定多名参与策划、指挥及实施人员的身份信息，其中包括：现任资通电军网络战联队网络环境研析中心负责人林钰书、现任资通电军网络战联队网络环境研析中心战队队长蔡杰宏、资通电军网络战联队网络环境研析中心现役人员粘孝帆及王浩铭。国家安全机关坚定捍卫国家安全，坚决打击台湾资通电军网络谋“独”、网络间谍活动，彻查幕后黑手，肃清危害隐患。



漏洞篇

知名前端开发工具 Vite 近期接连曝出 4 个任意文件读取漏洞。该系列漏洞是由于 Vite 开发服务器的文件访问控制机制存在缺陷且补丁修复不够全面，导致攻击者可以通过不同参数组合、路径构造或特殊字符屡次绕过 server.fs.deny 限制，非法访问项目根目录外的敏感文件。Vite 广泛应用于 Vue.js 项目开发，在国内有十万级的暴露风险资产，建议用户尽快做好自查及防护。

Foxmail for Windows 远程代码执行漏洞安全风险通告

4 月 11 日，奇安信威胁情报中心红雨滴团队于 2025 年年初在威胁情报狩猎过程中观测到客户网络中的异常行为，协助应急响应时溯源到最初的邮件攻击来源，提取到了相关邮件，分析显示攻击者组合利用 Foxmail 客户端存在的高危漏洞 (QVD-2025-13936)，受害者仅需点击邮件本身即可触发远程命令执行，最终执行落地的木马。情报中心第一时间复现确认了所发现的新漏洞，并将其上报给腾讯 Foxmail 业务团队。目前该漏洞已经被修复，最新版的 Foxmail 7.2.25 (2025-03-28) 不受影响，Foxmail 团队给予了致谢。

Vite 任意文件读取漏洞 (CVE-2025-32395) 安全风险通告

4 月 11 日，奇安信 CERT 监测到官方修复 Vite 任意文件读取漏洞 (CVE-2025-32395)，该漏洞源于 Vite 开发服务器在处理特定 url 请求时，没有对请求的路径进行严格的安全检查和限制，导致攻击者可以绕过保护机制，非法访问项目根目录外的敏感文件。奇安信鹰图资产测绘平台数据显示，该漏洞关联的国内风险资产总数为 47,506 个，关联 IP 总数为 10,961 个。目前该漏洞技术细节与 PoC 已在互联网上公开，奇安信威胁情报中心安全研究员已成功复现。鉴于该漏洞影响范围较大，建议客户尽快做好自查及防护。

Vite 任意文件读取漏洞 (CVE-2025-31486) 安全风险通告

4 月 7 日，奇安信 CERT 监测到官方修复 Vite 任意文件读取漏洞 (CVE-2025-31486)，该漏洞源于 Vite 开发服务器在处理特定 url 请求时，没有对请求的路径进行严格的安全检查和限制，导致攻击者可以绕过保护机制，非法访问项目根目录外的敏感文件。奇安信鹰图资产测绘平台数据显示，该漏洞关联的国内风险资产总数为 47,363 个，关联 IP 总数为 10,948 个。目前该漏洞技术细节与 PoC 已在互联网上公开，奇安信威胁情报中心安全研究员已成功复现。鉴于该漏洞影响范围较大，建议客户尽快做好自查及防护。

Zabbix groupBy SQL 注入漏洞安全风险通告

4 月 3 日，奇安信 CERT 监测到官方修复 Zabbix groupBy SQL 注入漏洞 (CVE-2024-36465)，该漏洞源于 include/classes/api/CApiService.php 文件对参数校验不当，导致具有 API 访问权限的 Zabbix 用户可以通过 groupBy 参数执行任意 SQL 命令。奇安信鹰图资产测绘平台数据显示，该漏洞关联的国内风险资产总数为 21,139 个，关联 IP 总数为 6156 个。目前，奇安信威胁情报中心安全研究员已成功复现。鉴于此漏洞影响范围较大，建议客户尽快做好自查及防护。

Vite 任意文件读取漏洞 (CVE-2025-31125) 安全风险通告

4 月 1 日，奇安信 CERT 监测到官方修复 Vite 任意文件读取漏洞 (CVE-2025-31125)，该漏洞源于 Vite 开发服务器在处理特定 url 请求时，没有对请求的路径进行严格的安全检查和限制，导致攻击者可以绕过保护机制，非法访问项目根目录外的敏感文件。奇安信鹰图资产测绘平台数据显示，该漏洞关联的国内风险资产总数为 47,578 个，关联 IP 总数为 10,830 个。目前该漏洞技术细节与 PoC 已在互联网上公开，奇安信威胁情报中心安全研究员已成功复现。鉴于该漏洞影响范围较大，建议客户尽快做好自查及防护。

CrushFTP 身份验证绕过漏洞安全风险通告

3 月 31 日，奇安信 CERT 监测到官方修复 CrushFTP 身份验证绕过漏洞 (CVE-2025-2825)，该漏洞源于处理身份验证标头不当，攻击者可绕过认证机制获取管理员权限，进而可能获取敏感信息、篡改数据或执行其他恶意操作。奇安信鹰图资产测绘平台数据显示，该漏洞关联的全球风险资产总数为 19882 个，关联 IP 总数为 6101 个。目前该漏洞技术细节与 PoC 已在互联网上公开，奇安信威胁情报中心安全研究员已成功复现。鉴于该漏洞影响范围较大，建议客户尽快做好自查及防护。

Ingress NGINX Controller 远程代码执行漏洞安全风险通告

3 月 26 日，奇安信 CERT 监测到官方修复 Ingress NGINX Controller 远程代码执行漏洞 (CVE-2025-1974)，该漏洞源于 Ingress NGINX Controller 没有充分验证 ingress 配置，攻击者可伪造 AdmissionReview 请求加载恶意共享库执行任意代码，访问 Kubernetes 集群敏感信息等。奇安信鹰图资产测绘平台数据显示，该漏洞关联的国内风险资产总数为 1601 个，关联 IP 总数为 1544 个。目前该漏洞技术细节与 PoC 已在互联网上公开，奇安信威胁情报中心安全研究员已成功复现。鉴于该漏洞影响范围较大，建议客户尽快做好自查及防护。

Vite 任意文件读取漏洞 (CVE-2025-30208) 安全风险通告

3 月 26 日，奇安信 CERT 监测到官方修复 Vite 任意文件读取漏洞 (CVE-2025-30208)，该漏洞源于 Vite 开发服务器在处理特定 url 请求时，没有对请求的路径进行严格的安全检查和限制，导致攻击者可以绕过保护机制，非法访问项目根目录外的敏感文件。奇安信鹰图资产测绘平台数据显示，该漏洞关联的国内风险资产总数为 47,173 个，关联 IP 总数为 10,657 个。目前该漏洞技术细节与 PoC 已在互联网上公开，奇安信威胁情报中心安全研究员已成功复现。鉴于该漏洞影响范围较大，建议客户尽快做好自查及防护。

Google Chrome 沙箱逃逸漏洞安全风险通告

3 月 26 日，奇安信 CERT 监测到官方修复 Google Chrome 沙箱逃逸漏洞 (CVE-2025-2783)，该漏洞源于 Google Chrome 沙箱机制与 Windows 操作系统内核交互时的逻辑错误，攻击者可利用该漏洞绕过沙箱隔离机制，造成信息泄露、代码执行等危害。目前该漏洞已发现在野利用。鉴于此漏洞影响范围较大，建议客户尽快做好自查及防护。

Windows 文件资源管理器欺骗漏洞安全风险通告

3 月 19 日，奇安信 CERT 监测到微软发布 3 月补丁日安全更新修复 Windows 文件资源管理器欺骗漏洞 (CVE-2025-24071)，该漏洞产生的原因是 Windows 资源管理器在解压包含特制 .library-ms 文件的 RAR/ZIP 存档时，会自动解析该文件内嵌的恶意 SMB 路径（如 \\192.168.1.116\shared），触发隐式 NTLM 认证握手，导致用户 NTLMv2 哈希泄露。目前该漏洞技术细节与 PoC 已在互联网上公开，奇安信 CERT 已成功复现。鉴于该漏洞已发现在野利用行为，建议客户尽快做好自查及防护。

AI大模型部署安全防护

实战指南

自 2025 年年初以来，人工智能的发展速度令人震撼。企业正处于要么部署 AI，要么被市场淘汰的重要的十字路口。但广泛采用 AI 大模型仍面临诸多挑战，包括安全隐患、幻觉等问题。 专题将主要介绍部署 AI 大模型的主要风险，以及如何构建数据安全防护机制、AI 预防性安全与治理控制措施、AI 运行时安全防护。



安全从业者和领导者的 AI 风险与安全防护指南

自 2025 年年初以来，人工智能的发展速度令人震撼。企业在广泛采用 AI 时仍面临诸多挑战，包括成本、幻觉及安全隐患等问题。

目前，企业正处于重要的十字路口：要么部署 AI，要么被市场淘汰。 网络安全领域也正迎来一个重大机遇——通过保障 AI 安全，帮助企业顺利落地 AI 方案，从而实现实质性商业成果。

本报告将提供详尽的指导，助力企业在 2025 年安全高效地部署 AI。

一、概要

1. 概述

2024 年，Software Analyst Cyber Research（SACR）发布了一份关于企业如何保障 AI 安全的深度研究报告。本报告是上一份研究的 2.0 版本。报告结合了过去 8 个月的市场洞察，分析了多个 AI 安全供应商，可帮助企业实现 AI 治理、数据安全和模型保护。

报告帮助读者深入了解以下关键内容：

a） 2025 年 AI 采用及相关安全风险现状。

b） 2025 年 AI 安全市场分类及供应商核心关注点。

c） 基于 CISO（首席信息安全官）访谈的 AI 安全部署建议。

2. AI 安全解决方案

研究发现，市场上许多供应商的能力存在较大重叠。一些供应商功能较为全面，而另一些则是相对较新的市场参与者。整体来看，三大核心能力尤为突出。

a） AI 治理控制（AI Governance Controls）

b） AI 运行时安全（Runtime Security for AI）

c） 生成式 AI 渗透测试与红队演练（GenAI Red Teaming & Penetration Testing）

3. 企业 AI 安全建议

a） 数据安全防护机制：企业在部署 AI 之前，必须优先考虑基础的数据安全防护措施。

b） AI 预防性安全与治理控制措施：企业应部署 AI 发现与清单管理解决方案，追踪 企业内的 AI 位置、用途及责任人。

c） AI 运行时安全防护：这是当前企业 AI 安全中改进空间最大的领域。本文详细分析了现有网络安全措施在应对



人工智能 AI 安全市场分类

AI 特定风险时的缺陷与局限性。

二、简介

人工智能为企业带来了重大变化和机遇。2023 年，人工智能兴起，一些企业成为早期采用者。2024 年，企业采用人工智能的数量显著增加，但此前观察到结果显示无法与 2025 年的浪潮相比。由于安全和隐私风险问题，许多公司一直犹豫是否采用人工智能。然而，到 2025 年，未采用人工智能的企业将不能再继续保持竞争力。现在到了企业必须积极采用人工智能的时候了。

· **采用 AI 的收益。**采用 AI 可以节省成本和开支。对于许多企业来说，采用人工智能的投资回报主要是成本节约，而不是新的收入机会。预计到 2025 年，企业在人工智能上的支出将增加约 5%。IBM 公司表示，该公司计划将其收入的平均 3.32% 分配给人工智能——对于一家价值 10 亿美元的公司来说，这相当于每年 3320 万美元。

· **采用 AI 的挑战。**采用 AI 可以带来巨大的好处，但也带来了重大的挑战，无论是 LLM、GenAI 还是 AI 智能体，CISO、CIO、数据副总裁和工程副总裁均担忧安全、隐私、偏见和合规问题。人工智能部署失误已导致多个企业遭受财务损失和声誉受损。例如，加拿大航空的聊天机器人向客户提供了丧亲票价的误导性信息，被勒令向客户退款。另一起案件中，一家韩国初创公司在聊天中泄露了敏感客户数据，被韩国政府罚款约 9.3 万美元。损失金额看起来像是“轻微”罚款，但情况可能要糟糕得多。2023 年 2 月，谷歌在其 Bard AI 聊天机器人分享了不准确信息后，市值缩水近了 1000 亿美元。

· **知名 CISO 观点：**

“这是安全首次成为业务驱动力——安全积极支持业务发展。人工智能 (AI) 和大型语言模型 (LLM) 正在改变行业、提高效率并释放新的商机。快速采用带来了重大的安全、合规和治理挑战。组织必须在 AI 创新与风险管理之间取得平衡，确保监管一致性、客户信任和董事会层面的监督。如果没有结构化的安全策略，公司将面临敏感数据泄露、对手操纵和利益相关者信心受损的风险。AI 安全不再仅仅是运营问题，而是一项战略要务，直接影响企业风险、监管风险和长期业务可行性。”

三、2025 年企业 AI 大模型部署现状

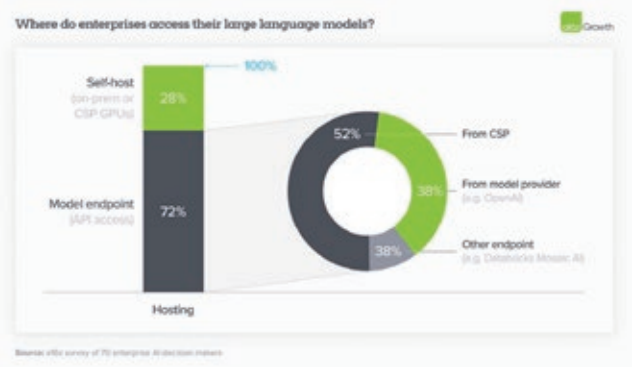
1、2025 人工智能部署率继续攀升

正如本报告前文所述，企业对 AI 的投资市场正在快速增长。《全球生成式 AI 安全就绪度报告》指出，42% 的企业已在多个业务领域积极部署大语言模型（LLM），另有 45% 的企业正在探索 AI 应用。AI 的采用正在多个行业加速发展，尤其是在法律、娱乐、医疗和金融服务领域增长显著。

根据 Menlo Ventures《2024 企业生成式 AI 现状报告》，行业的 AI 相关支出从 1 亿到 5 亿美元不等。报告还指出，企业级生成式 AI 的主要应用场景还包括代码生成、搜索与检索、摘要提取以及智能聊天机器人。

整体来看，企业 AI 采用率持续攀升，目前超过 85% 的企业已在云环境中使用托管或自托管的 AI 解决方案，表明 AI 采用率虽然趋于稳定，但仍在增长。其中，托管 AI 服务的使用率已从前一年的 70% 增长至 74%，显示出企业对托管 AI 解决方案的持续青睐。





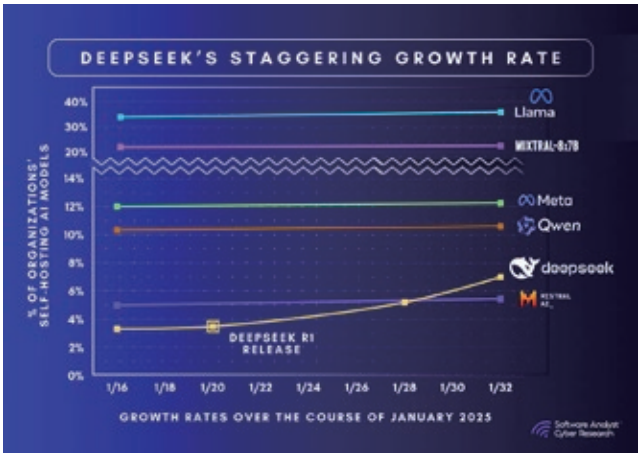
2、开源 vs. 闭源模型：未来将趋平衡

企业最初主要使用闭源模型，但如今越来越多地转向开源模型。Andreessen Horowitz 的《企业构建和购买生成

式人工智能方式的 16 项变化》的报告预测，41% 的受访企业将在其业务中增加使用开源模型来代替闭源模型。该报告还预测，如果开源模型的性能与闭源模型相当，另外 41% 的企业将从闭源模型转向开放模型。

开源模型的优势包括由于无许可费用而降低成本、全球社区快速创新、定制化和透明度，这有可能推动企业更多地采用。许多最受欢迎的 AI 技术要么是开源，要么与开源生态系统紧密相关。

在 DeepSeek 引发颠覆性变革后，人们对开源模型的兴趣进一步升温。尽管 DeepSeek-r1 存在许多安全问题，但其性能表明开源模型在性能和成本方面都与闭源模型具有很强的竞争力。



Andreessen Horowitz 预测，未来开源与闭源 AI 模型的使用比例将趋于平衡。相较于 2023 年市场 80% ~ 90% 由闭源模型主导的情况，市场趋势发生了显著变化。



资料来源：State of AI Wiz Research

选择闭源模型的团队通常采用商业授权模式，支付相应费用，并签署协议，确保模型提供商不会将企业输入的数据用于训练。这种模式为企业提供了更强的数据隐私保障和合规性支持。

选择开源模型的团队则需要更加关注 AI 模型生命周期的安全，确保模型能够按照预期输出结果，并且具备较强的抗越狱能力，以防止被滥用或攻击。

3、托管 vs 自托管：自托管大幅跃升

许多企业都依赖托管 AI 产品（到 2025 年，这一比例将从 70% 增长到 74%），但报告强调，自托管 AI 的采用率大幅上升，同比从 42% 跃升至 75%。向自托管 AI 的转变需要强大的治理来保护云环境。

4、安全影响

企业在 AI 采购和部署方面面临诸多选择，这也涉及 AI 安全方案的采购。首席信息安全官（CISO）更倾向于选择能够在现有架构内运行、无需依赖外部第三方来维护数据隐私的解决方案。同时，CISO 和产品高管希望 AI 安全产品具备可配置性，使团队能够针对不同项目定制检测设置。例如，某些项目可能需要对 PII（个人可识别信息）泄露采取更严格的敏感性控制，而另一些项目则需要加强对有害内容的监控。

归根结底，AI 和模型安全是一个快速发展的领域。CISO、CIO 及产品和工程负责人普遍强调，AI 安全产品必须具备快速创新能力，以跟上行业发展步伐。首席信息安全官强调，真正的 AI 威胁可能在未来 2~5 年内才会完全显现，因此，AI 安全产品必须持续应对新兴威胁，以确保长期安全性。

四、企业 AI 部署中的风险与挑战

根据我们的研究，进入 2025 年后，企业面临的风险依然严峻，尤其是在 DeepSeek 等开源模型出现显著进展后。在众多安全框架中，以下三个框架是企业最常用的 AI 风险分类与评估工具。

1、AI 威胁治理框架

· **OWASP 大模型和生成式 AI 威胁框架**：CISO 依靠 OWASP 十大威胁框架来识别和缓解关键的 AI 威胁。该框架重点介绍了诸如提示词注入（#1）和过度代理（#6）等关键风险，帮助团队实施适当的输入验证并限制 LLM 权限。

· **MITRE ATLAS**：MITRE 提出的人工智能系统对抗性威胁全景（ATLAS）框架全面梳理了 AI 系统可能面临的威胁与攻击手法。该框架概述了服务拒绝和模型中毒等场景。其他关键资源包括 Google 安全 AI 框架、NIST AI 风险管理框架和 Databricks 的 AI 安全框架等。

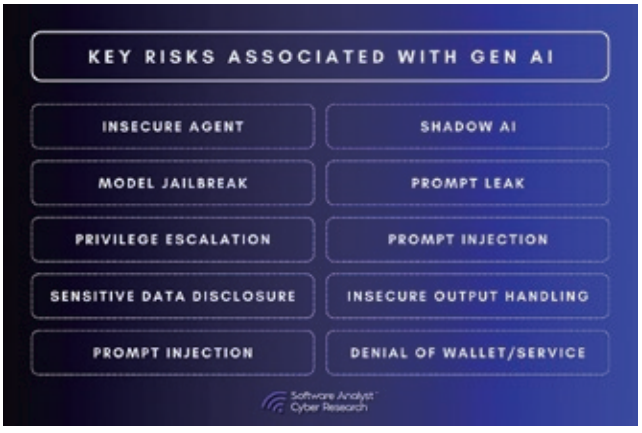
· **生成式 AI 攻击矩阵**：该知识源矩阵系统化梳理了攻击者针对生成式 AI 系统、智能助手（Copilot）及智能体所采用的战术、技术与流程（TTP）。其设计灵感源自 MITRE ATT&CK 等框架，旨在帮助组织识别 Microsoft 365、容器及 SaaS 环境中特有的安全风险。

企业面临的 AI 风险可以分为两个主要领域：

1） 员工行为与 AI 使用安全防护。企业需要确保员工在使用 AI 工具（如 ChatGPT、Claude、Copilot 等）时不会无意间泄露敏感信息或违反合规规定。主要风险包括： 数据泄露与合规风险、影子 AI 风险、风险 AI 生成内容（AIGC）、访问控制与权限管理。

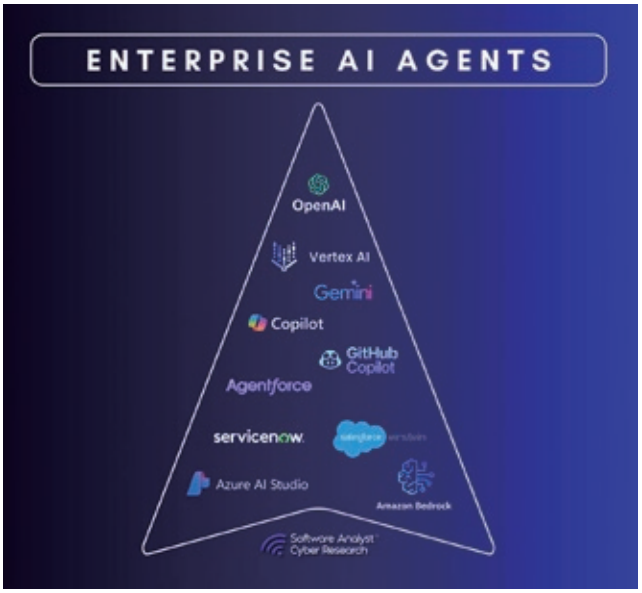
2） 自研 AI 应用的全生命周期安全防护。对于开发和部署自研 AI 应用的企业来说，AI 模型的安全性、完整性和运行时防护至关重要。主要风险包括：AI 供应链安全、运行时攻击与 AI 保护、生成式 AI 风险管理。

2、围绕 AI 使用、活动和防护措施的安全风险



最近数月，集成 Microsoft Copilot、Google Gemini、ServiceNow 和 Salesforce 等 AI 工具的企业遇到重大安全挑战。

主要问题是权限过高，AI 助手访问的数据超出了必要范围，导致敏感信息意外泄露。例如，Microsoft Copilot 与 Microsoft 365 的深度集成使其能够聚合大量组织数据。如果权限管理不善，可能会造成安全漏洞。



以下是截至 2025 年发现的突出风险，主要集中在人工智能的使用、数据暴露和政策违规方面。

· **数据泄露和隐私问题**：企业希望更好地了解未授权 AI 系统相关的数据泄露风险。这对于金融和医疗保健等受监管行业的机构尤其重要，因为数据泄露可能会导致严重的监管处罚和声誉损害。2025 年，通过 AI 应用处理敏感数据的企业正在积极评估和实施有力的保障措施，以确保遵守 HIPAA、GDPR 和其他法规。

· **影子 AI 泛滥**：2025 年企业面临的一个主要挑战是员工广泛采用未授权的 AI 工具。这些未经批准的人工智能应用（包括广泛使用的聊天机器人和生产力工具）在组织控制之外运行，构成了重大的安全风险。市场已经出现了大模型防火墙和其他保护措施来解决这个问题，但迅速扩张的人工智能领域使全面监控变得越来越复杂。企业必须通过明确定

义的人工智能使用政策和技术控制来平衡创新与安全。

五、保障自研 AI 应用全生命周期安全

1、AI 全生命周期安全保障不同于传统软件开发 (SDLC)

确保 AI 开发生命周期的安全面临着与传统软件开发不同的独特挑战。虽然、静态应用安全测试 (SAST)、动态应用安全测试 (DAST) 和 API 安全等传统安全措施仍然适用，但 AI 开发与部署流程（AI Pipeline）带来了额外的复杂性和潜在的盲点。

· **扩大攻击面**：数据科学家经常使用 Jupyter 笔记本或 MLOps 平台等环境，这些环境在传统的持续集成 / 持续部署 (CI/CD) 管道之外运行，从而产生潜在的安全盲点，往往比传统代码更隐蔽。人工智能系统面临着特定的威胁，包括数据投毒和对抗性攻击，恶意攻击者将损坏或误导性的数据引入训练数据集，从而损害模型的完整性。

· **复杂的 AI 全生命周期组件**：AI 开发过程包含不同的阶段——数据准备、模型训练、部署和运行时监控。每个阶段都会引入独特的漏洞和潜在的错误配置，因此需要专门的安全检查点。

· **供应链安全和 AI 物料清单 (AI-BOM)**：AI 应用通常依赖大量数据，可能会使用开源模型或数据集。如果没有经过细致的审查，这些组件可能会引入漏洞，导致 AI 系统被破坏。鉴于复杂的依赖关系（包括脚本、笔记本、数据管道和预训练模型），对 AI 物料清单 (AI-BOM) 的需求日益增加。AI 物料清单（AI-BOM）是一份完整的清单，有助于识别和管理 AI 供应链中的潜在安全风险。

在不深入探讨的情况下，其他包括监控 RAG 访问、治理与合规问题及运行时安全需求。

2、AI 开发生命周期分步说明

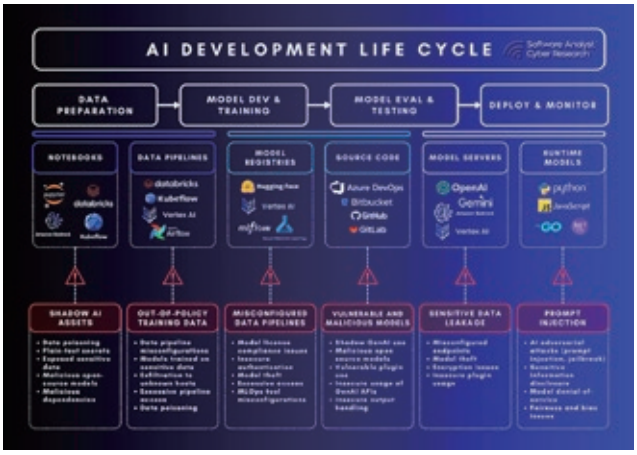
人工智能（AI）开发生命周期引入了独特的安全挑战，需谨慎考量。明确构建和开发 AI 模型的核心步骤至关重要。

· **数据收集和准备：**收集、清理和预处理数据，确保数据质量和合规性。在应用特征工程（Feature Engineering）时将数据拆分为训练集、验证集和测试集。

· **模型开发与训练**：选择合适的模型类型和框架，利用优化技术进行训练并进行微调，以提高准确性。解决偏见、公平性和对抗性攻击等安全风险。

· **模型评估与测试**：使用关键指标评估性能，进行对抗性测试，对极端情况进行压力测试，并确保可解释性。检测数据中毒、后门和模型漂移。

· **部署和持续监控**：通过 API、云端或边缘设备部署模型，实施实时安全监控，通过 MLOps 自动重新训练并审核 AI 决策，以确保合规性和安全性。



3、开发 AI 应用各阶段相关的风险

· **数据丢失**：企业在开发 AI 应用时面临重大的安全风险。其中数据泄露是最令人担忧的问题之一。特别是在使用连接互联网的 AI 模型或缺乏合同保障防止模型使用客户数据训练时，数据泄露的可能性更高。这些风险不是理论上的——它们现在正在发生。企业必须彻底检查模型提供商的隐私政策和数据处理实践。

· **易受攻击的模型**：MLSecOps 的崛起与 AI 安全防护。随着企业逐步认识到 AI 开发流程不同于传统软件开发，MLSecOps（机器学习安全运营）正在迅速兴起。AI 开发人员（如机器学习工程师和数据科学家）通常在 Databricks、Jupyter Notebooks 等专门环境中工作，而非传统的软件开发流水线。这些非标准化的开发环境需要严格监控，以防止错误配置、API 密钥泄露、数据外泄等安全风险。类似于软件供应链安全扫描，MLSecOps 需要

定期扫描 AI 模型，以发现潜在的后门、漏洞和恶意修改。企业需要严格审查模型来源，防止供应链攻击。因为攻击者可以从开源存储库下载 AI 模型、恶意修改，使其可将数据传输到黑客服务器，并重新发布让毫无戒心的用户下载使用。

· **数据投毒**：数据投毒（Data Poisoning）是 AI 安全中的关键风险，它可能导致 AI 模型被植入后门，进而影响 AI 预测结果，甚至让模型执行恶意行为。Orca Security 发布了一款名为 AI Goat 的 AI 安全教育工具，其中专门演示了数据投毒攻击的危害。其核心原理是，攻击者上传新数据或篡改 AI 训练数据，让模型在学习过程中受到污染，从而改变其行为。

数据投毒与模型投毒密切相关，这是一种允许攻击者在模型中插入后门的威胁。后门植入后，模型在普通输入下表现正常，但在特定触发条件下会执行恶意操作。传统安全扫描工具难以发现 AI 模型中隐藏的后门。2024 年，人类学研究人员在一篇题为《潜伏特工：训练通过安全训练持续存在的欺骗性 LLM 》的论文中揭示了这一点。

· **安全团队缺乏机器学习专业知识**：

通过机器学习引入的欺骗性行为已成为 AI 安全领域的重要关注点。 OWASP AI/LLM 安全研究员 Emmanuel Guilherme 在《全球生成式 AI 安全就绪度报告》中谈到了机器学习的复杂性及其对人工智能安全的影响。Guilherme 表示：“保护 AI 系统的最大障碍是可见性缺失，尤其是在使用第三方供应商时。理解机器学习流程（ML flow）的复杂性，以及对抗性机器学习（Adversarial ML）的微妙之处，使这一挑战更加严峻。 构建强大的跨职能机器学习安全团队极具挑战性，需要来自不同背景的专业人员共同创建全面的安全场景。”机器学习在 AI 应用开发中扮演着重要角色，这一领域的复杂性催生了一个新的术语——MLSecOps。MLSecOps 专注于在机器学习工作流（如数据管道和 Jupyter Notebooks）中嵌入安全机制，确保 AI 开发过程中的安全性。

· **提示词攻击和越狱**：在 AWS re:Inforce 2024 大会上，亚马逊首席安全官 Stephen Schmidt 强调，AI 应用需要持续测试，因为大语言模型不仅在发布新版本时会发生变化，还会随着用户的交互不断演变传统的安全方法，只关注左移扫描——即在代码发布前进行评估，但这种方式无法满足 AI 应用的安全需求。AI 应用需要持续的安全扫描，尤其是在运行时（Runtime）。在运行时，AI 应用可能会遇到提示词攻击，如直接提示词注入、间接提升词注入和越狱。这些攻击可能导致未经授权的访问、数据泄露或生成有害输出。

六、现有网络安全控制措施

目前，企业主要采用三种方法来保护 AI 系统。没有一种万能的方法。通常，企业会结合其中 2~3 种方法来增强 AI 的安全性。

- 模型提供者和自托管解决方案。
- 现有的安全控制措施和供应商。
- 引入纯 AI 安全供应商。

1、模型提供商和自托管解决方案

此类供应商包括 OpenAI、Mistral 和 Meta 等 AI 公司。这些企业在 AI 模型的训练和推理过程中，将安全性、数据隐私和负责任的 AI 保护嵌入到他们的模型中。他们的主要重点仍然是构建最先进、最高效的 AI 模型，而不是优先考虑安全性。因此，许多企业寻求独立的安全层，以保护在不同环境下运行的所有 AI 模型。

- 自托管模型（Self-hosted Models）：部分企业选择自托管 AI 模型，主要是为了增强安全性和数据隐私保护，确保敏感信息不会离开企业内部可控的基础设施。这种方法在金融、医疗和政府等受严格监管的行业尤为重要。通过自托管，企业可以完全控制模型的更新、微调和安全协议，减少对外部供应商的依赖，避免可能的安全漏洞。
- 第三方模式（Third-party Models）：使用第三方 AI 模型存在一定的安全风险，因为数据需要传输到外部服务器，从而增加数据泄露、政策变更和供应商锁定（Vendor Lock-in）的风险。因此，完全依赖第三方云端 AI 模型的企业通常会部署额外的独立安全措施，以减少潜在的威胁。

2、传统网络安全控制措施

许多企业已经部署了一定程度的网络安全措施，以保护数据、网络和终端设备。研究发现，缺乏专门 AI 安全解决

方案的企业通常依赖现有控制措施。然而，这些措施在 AI 安全领域各有优劣，企业需要充分了解其能力和局限性，以制定有效的 AI 安全策略。



1） 网络安全

· **防火墙 /SASE**：传统网络安全措施（包括防火墙和入侵检测系统）有助于保护 AI API 端点及数据流量，防止未经授权的访问和横向移动攻击。部分企业会实施专门针对 AI 的防火墙规则来限制高风险 API 查询，例如，限制 LLM API 请求的速率以防止模型提取。但传统防火墙难以检测恶意 AI 提示词注入（Prompt Injection）或防范 AI 模型篡改（如后门攻击）。一些供应商（如 Witness AI）已经与 Palo Alto Networks 等网络安全公司合作，强化 AI 相关策略执行。

· **Web 应用防火墙 (WAF)**。WAF 也可以限制 AI 推理请求的速率，以防止模型抓取（Model Scraping），并缓解自动提示词注入攻击（Prompt Injection）。但 WAF 并不具备专门的 AI 威胁检测能力。

· **云访问安全代理 (CASB)**。CASB 可查看和控制流经云环境的 AI 相关数据，从而可有效保护 AI SaaS 应用并防止影子 AI 部署。但 CASB 主要关注访问控制和云数据安全，并不直接保护 AI 模型本身。

2） 数据丢失防护 (DLP) 和数据安全态势管理 (DSPM)

数据丢失防护 (DLP) 和数据安全态势管理 (DSPM) 解决方案旨在保护敏感信息免遭未经授权的访问或泄露。在 AI 系统中，这些工具有助于防止意外或故意泄露个人身份信息 (PII) 或专有业务数据。然而，它们往往无法解决特定于 AI 的威胁，例如，通过 API 抓取或分析 AI 模型逻辑的复杂性而导致的模型泄露。此外，如果没有 AI 感知策略，这些工具可能无法有效缓解大型语言模型 (LLM) 特有的风险，如提示泄露（Prompt Leakage）。

3） 云安全 (CNAPP 和 CSPM)：由于 AI 模型经常部署在 AWS、GCP 或 Azure 等云平台上，云原生应用保护平台 (CNAPP) 和云安全态势管理 (CSPM) 工具可以识别 AI 模型部署中的错误配置和安全漏洞。许多供应商已经开始扩展相关功能，以专门检测云端 AI 模型的漏洞。

4） 端点检测和响应 (EDR)：终端检测与响应（EDR）解决方案适用于检测恶意软件和传统端点威胁，但通常无法监控 AI 相关风险，如未经授权的推理尝试（Unauthorized Inference Attempts）或 API 抓取攻击（用于提取模型功能）。

5） 应用安全测试 (SAST/DAST)：静态和动态应用安全测试工具能有效评估传统应用程序的漏洞。然而，它们很难评估 AI 模型行为或评估其抵御对抗性攻击（Adversarial Attacks）的能力。目前，一些新兴供应商正在尝试弥补这一缺陷。

6） 浏览器安全：浏览器安全措施可以控制 AI 聊天机器人的使用，但它们无法保护专有 AI 模型本身。这些控制仅限于用户交互层面，并未扩展到底层 AI 基础设施。

许多传统网络安全供应商已经开始检测 AI 相关流量，并尝试覆盖 AI 安全领域。一些供应商开始将 AI 安全防护功能纳入其产品中。然而，该领域的创新仍然缓慢，这些供应商尚未提供突破性的 AI 安全功能。此外，由于传统安全供应商涵盖了广泛的安全领域（包括云、网络和端点安全），AI 安全通常只是其产品组合的一个方面，而不是重点。

3、新兴专攻型解决方案

第三类包括专注于 AI 安全的新兴企业。这些公司大多聚焦于细分领域的挑战，包括保障 AI 使用安全，防御提示词注入攻击（Prompt Injection Attacks），识别开源大语言模型（LLMs）中的漏洞。从当前分析可见，AI 安全需求迫

切且覆盖领域广泛，其问题集合庞杂多样，需在快速演进的生态中持续应对。

七、市场主要 AI 安全防护方案

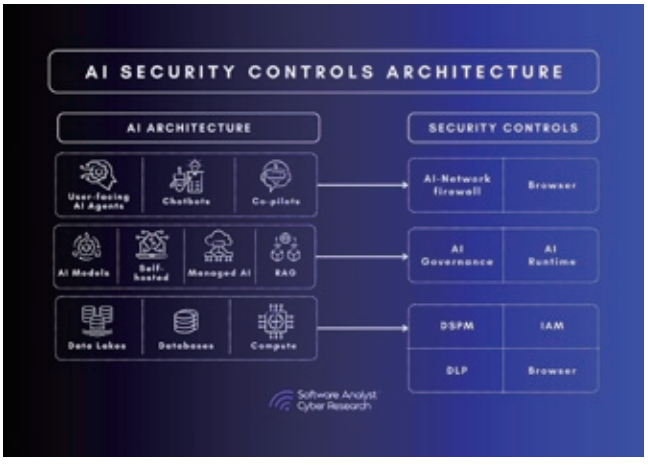
本部分基于市场研究为企业提供现有市场动态指南。

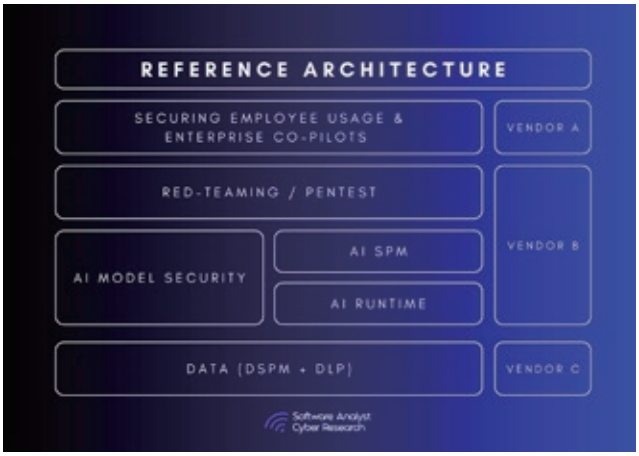
1、数据与计算层（AI 基础层）。企业必须保障数据库、数据湖或云存储服务（如 AWS S3）数据的安全。有些企业可能已经拥有矢量数据库。企业应实施强大的数据治理控制措施。

2、AI 模型与推理层（确保 AI 逻辑安全）。企业通常会组合使用基础模型、微调模型（开放源或闭源）、嵌入式数据、推理端点以及智能 AI 工作流。这些模型可能托管在内部或公共云中。企业必须保障整个 AI 生命周期的安全，从构建到运行时。

3、AI 应用和用户界面层（确保 AI 使用安全）：企业通常会采用 AI 助手（Copilot）、自主智能体（Autonomous Agents）或聊天机器人（Chatbots）。企业必须保护提示词安全（Secure Prompting）、实施访问控制（RBAC）及内容过滤（Content Filtering），防止 AI 生成不当内容或导致数据泄露。

在上述各个层级中，企业应根据自身的 AI 战略，在数据与计算层、AI 模型与推理层、AI 应用与用户界面层选择一到两家专业安全供应商，以全面保护 AI 生态系统。





八、针对安全 AI 部署安全的建议

- 数据安全控制
- AI 预防性安全和风险控制
- AI 运行时安全

1、AI 数据安全控制

强大的 AI 安全策略离不开稳健的数据安全控制，因为 AI 模型的安全性和可信度取决于其训练和交互的数据。AI 系统引入了与数据访问、完整性、泄漏和治理相关的新风险，这使得数据安全成为 AI 部署的基础要素。在《全球生成式 AI 安全准备报告》中，受访者（包括首席信息安全官、业务用户、安全分析师和开发者）将数据隐私和安全视为 AI 采用的主要障碍。Vanta 的信息与合规经理 Adam Duman 指出，人工智能会让现有的安全问题更加严重。如果公司正在构建 GenAI 应用或 AI 代理，他们希望使用自己的数据，并会考虑可信度、负责任的输出和避免偏见。例如，如果一家公司没有现有的数据治理措施，如数据标记政策，这些问题在 AI 中会变得更加严重。

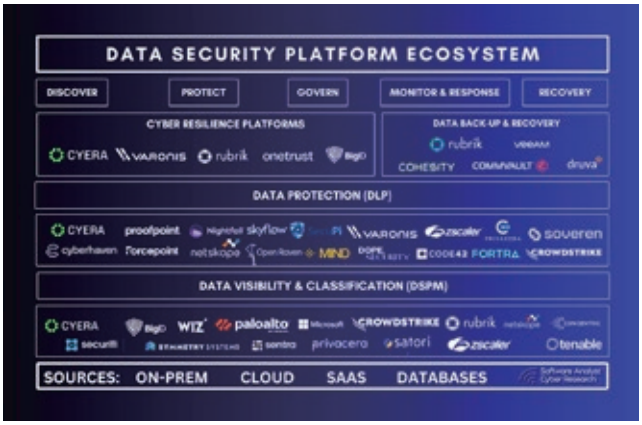
企业在选择安全供应商时，应重点关注如何在 AI 利用数据之前的每个阶段确保数据安全。关键措施包括以下。

1） 数据安全态势管理（DSPM）： DSPM 解决方案可扫描企业环境，提供全面的数据可见性，并帮助企业识别需要符合 GDPR、CCPA、HIPAA 及 AI 相关法规的数据集，确保 AI 部署合法且合规。原因是 AI 系统需要大量结构化和

非结构化数据，但并非所有数据都应该可供 AI 模型访问。因此，公司应该使用 DSPM 来发现敏感数据。

2） 角色与属性访问控制（RBAC & ABAC）：企业应部署基于角色（RBAC）和基于属性（ABAC）的访问控制，以确保 AI 模型和员工只能访问授权数据集。

3） 数据丢失预防（DLP）：公司应防止敏感数据泄露到生成式 AI 模型中，因为员工可能会无意中输入受监管或机密的信息。



2、AI 预防与治理控制措施

组织应采取主动措施保护人工智能。这些关键要素包括如下

· **AI 治理策略（Governance Policies）**：企业必须制定全面的治理和安全策略，以确保 AI 的安全部署和使用可控。治理控制应覆盖 AI 使用情况监测，确保企业内所有 AI 使用（无论是否正式批准）都受到监管； AI 应用开发与部署全生命周期安全保护。

· **AI 发现与资产清单管理**（哪里、什么和谁）：企业应部署 AI 资产发现和盘点管理解决方案，以回答“AI 在哪里、做什么、由谁使用？”企业无法保护他们看不到的东西——这一基本的网络安全原则同样适用于 AI。组织必须了解 AI 在内部的使用情况：员工是否与 ChatGPT、Claude 桌面版交互，或从 HuggingFace 下载开源模型？通过映射已批准和影子 AI 的使用情况，企业可以有效地扫描模型，生成 AI 物料清单（AI-BOM），以验证模型来源，并监控数据丢失。使用开源模型的企业应该扫描所有模型，分析不同 AI 模型的安全性与防护能力。通过 MLSecOps 了解模型来源有助于识别潜在的偏见。企业还应仔细审查供应商隐私政策，评估数据传输路径，确保数据主权合规，以确定组织数据的传输位置。

· **AI 安全态势管理**（SPM）解决方案：AI-SPM（人工智能安全态势管理）方案可对 AI/LLM 部署进行安全评估、风险监测和策略实施，从而提供可见性、风险检测、策略实施和合规性。AI-SPM 方案保护 AI 模型、API 和数据流，免受基于模型的攻击和数据泄露。利用云安全态势管理（CSPM）和数据安全态势管理（DSPM）功能，AI-SPM 可确保安全团队拥有实时的 AI 资产清单。AI-SPM 在整个 AI 生命周期中检测错误配置、数据泄露和模型完整性风险。此类别的 AI 安全公司可以帮助企业了解 AI 的使用位置和所连接的工具。

· **AI 渗透测试与 AI 红队**：Gen-AI 渗透测试和 AI 红队在识别漏洞、测试弹性和增强整个 AI 模型生命周期的安全性方面发挥着关键作用。它们的主要功能是模拟对抗性攻击，以在 AI 模型中发现安全漏洞，以免它们在现实场景中被利用。AI 渗透测试和红队应作为主动控制措施来实施，以确保安全的 AI 开发实践、政策合规性，以及在合规和治理层面进行风险评估。企业还应在实时环境中进行红队模拟，以检测模型漂移和运营风险。

为了总结本节内容，我们将使用首席信息安全官（CISO）的话。他将 AI 预防控制表述为：“为了应对这些风险，企业必须为 AI 集成零信任架构（ZTA），确保只有经过身份验证的用户、应用程序和工作流可以与 AI 系统交互（云安全联盟，2025 年）。AI 数据治理和合规框架还必须与欧盟 AI 法案和 SEC 网络规则保持一致，以提供透明度和董事会级别的 AI 风险管理可见性。通过嵌入 AI 安全设计，企业可以在保持业务敏捷性和创新的同时，满足监管要求。 对于大规模部署 AI 的企业，主动风险缓解至关重要。实施 AI 安全物料清单（AI-SBOMs）加强供应链安全，而 AI 红队测试则在被利用之前识别漏洞。随着 AI 法规的演变，安全领导者必须投资于持续监控、网络风险量化（CRQ），以及与行业风险管理框架（如 NIST、ISO/IEC、OECD）合规的一致性，以保持韧性。将 AI 安全嵌入其核心战略的组织——”

3、AI 运行时安全

企业必须为 AI 模型实施运行时安全，在推理和主动部署期间提供实时保护、监控和威胁响应。运行时安全应利用检测和响应代理、eBPF 或 SDK 进行实时保护。运行时是 AI 安全的重要组成部分。在题为“潜伏代理：在安全训练中持续存在的欺骗性 LLM 训练”的论文中，Anthropic 研究人员展示了带有后门的模型在训练和部署中如何故意表现出不同的行为。此类别的 AI 安全供应商可识别带有隐藏后门的模型，并监控传入的提示注入或越狱尝试。这些解决方案还可以检测重复的 API 请求，这些请求表明存在模型盗窃或其他可疑活动。供应商还应针对基于 API 的 AI 服务实施实时身份验证。

可观察性是运行时安全的延伸。目标是让企业捕获 AI 活动的实时日志和监控，如捕获推理请求、模型响应和系统日志以进行安全分析。它允许企业在实时中跟踪训练进度和 AI 响应。可观察性具有两个关键功能：一是帮助团队改进和调试 AI 聊天机器人；二是能够检测有害或意外的响应——在需要时允许系统抛出异常并提供替代响应。这一领域的 AI 安全供应商帮助企业有效地调试应用程序和评估响应。从构建到部署的整个生命周期需要持续监督。由于 AI 的非确定性，即使是强大的安全防护措施也不能保证一致的响应。这些解决方案应与 SIEM（安全信息和事件管理）和 SOAR（安全编排、自动化和响应）系统集成，以实现实时事件响应和分析。

九、行业现状与未来预测

AI 安全生态系统庞大且充满挑战，AI 安全供应商可划分为两个主要类别，供应商数量已超过 50 家。这种市场碎片化也带来了极大的管理挑战，尤其是对首席信息安全官（CISO）而言，可能成为一场噩梦。

- **评估与集成难度高**：企业需要分析、筛选、集成多个安全工具，可能导致安全架构臃肿，增加运营负担。
- **功能重叠导致选择困难**：不同供应商的 AI 安全方案功能存在大量重叠，导致企业在选择时难以区分最优方案。
- **安全工具集成不顺畅**：不同供应商的安全工具可能缺乏兼容性，增加企业的技术整合成本。
- **供应商长期稳定性存疑**：市场上新兴厂商大量涌现，但部分供应商的长期生存能力和支持能力尚不明朗，企业用户在技术能力之外，还需评估供应商的稳定性。

随着市场成熟，供应商整合（并购或关闭）可能会成为趋势，导致小型供应商被收购或倒闭，这可能影响企业已部署的安全方案。企业需要在专业化安全解决方案与供应商稳定性之间找到平衡，同时保持灵活的安全战略，以适应不断变化的市场环境。

预计成熟的网络安全公司将利用对 AI 安全解决方案日益增长的需求，但初创公司也仍有空间进行竞争。无论未来的创新 AI 安全解决方案来自成熟安全公司还是新兴企业，最终的赢家将是那些能够持续创新、主动适应新威胁的安全企业。

总的来说，AI 安全并非理论概念——现实中已经发生了从意外数据泄露到 AI 系统漏洞被利用等各种安全事件。幸运的是，在正确实施的情况下，AI 安全解决方案可以有效降低这些风险。同时，市场上也涌现了大量应对这些挑战的供应商和解决方案。

要有效防御这些威胁，企业必须实施贯穿 AI 全生命周期的安全措施——从数据收集、模型训练到部署和监控，确保每个环节都得到充分保护。

展望未来，随着企业不断分享经验，以及行业标准持续演进，AI 安全解决方案的有效性将进一步提升。市场正朝着一个方向发展：AI 在部署时就具备内置的安全性、隐私性和可信度，而不是事后补救。

关于作者

Software Analyst Cyber Research

专注于网络安全行业分析的机构，致力于为安全领域的领导者提供深入的行业动态分析和见解。



CTFWAR.ORG

网络安全的本质是攻防对抗
讲百遍不如打一遍

---习近平

CTFWAR介绍

CTFWAR攻防战争平台是CTFWAR网络安全攻防对抗联赛的官方平台，是由中国网络安全攻防大咖联合发起的创新型学习平台，以游戏的形式融入多种网络攻防场景进行答题、竞赛、互动。CTFWAR攻防靶场分为初级新手区、中级进阶区、高级挑战区，同学们可根据自身技术能力及技术方向进行筛选，整个学习过程将有全面的数据化呈现。

攻防答题模式

CTFWAR攻防战争平台采取主流的CTF夺旗赛和AWD攻防赛的答题模式，提供云端的攻击终端，同学们在云端攻击平台上进行攻防实战操作，通过 Flag 判定和操作用时进行综合评分，以获得积分升级和金币奖励。

积分等级体系

CTFWAR攻防战争平台采取主流的CTF夺旗赛和AWD攻防赛的答题模式，提供云端的攻击终端，同学们在云端攻击平台上进行攻防实战操作，通过 Flag 判定和操作用时进行综合评分，以获得积分升级和金币奖励。

红帽人才工程

Cyber Crime Governance Talent Training Project

工程简介

在“全国网络警察培训基地”的指导下，中国下一代网络安全联盟牵头发起「红帽人才工程」，联合华云信安、美亚柏科、高联通信等知名网络安全企业，围绕“网络犯罪治理、涉网犯罪打击”等相关网络安全课题，持续挖掘、培养、资助、赋能网络安全人才，构建红色基因的网络空间安全人才防线。

申报流程

课题征集 → 课题公示 → 立项评审

研究课题计划征集

研究课题信息公示

研究课题立项评审

申报说明

项目资讯

培养对象

政企单位在职人员、高校全日制在校生（含研究生）、互联网企业技术人员、网络安全企业技术人员以及社会网络安全人才...

核心课题

内网攻防、社工钓鱼、远控木马、免杀加壳、情报甄别、资金分析、调证溯源、取证固证、远程勘验、APP反编译、二进制逆向...

首度揭秘：奇安信 AI 大模型应用如何防范风险？

“忽如一夜春风来，千树万树梨花开。”自从 ChatGPT 发布之后，大模型以迅雷不及掩耳之势，掀起了向各行业领域渗透普及的浪潮。作为网络安全行业的领军企业，奇安信也不例外。

清华 GLM、零一万物、百川智能、LLaMA、通义千问、DeepSeek……仅仅一年多的时间，各种主流大模型在奇安信陆续部署。现在，奇安信集团依托自研安全大模型不仅面向公众提供服务，如安全知识问答平台；还开发了一系列产品服务，如 AISOC、AI+ 代码卫士、AI 终端安全、安全机器人等；同时，基于大模型的代码助手已经成为公司开发人员不可或缺的提效工具。可以说，大模型已经渗透到奇安信集团研发、生产、管理、运营、服务等各个环节。

奇安信集团副总裁、研发部负责人张勇表示，“AI 大模型技术的发展速度之快超出想象；奇安信集团的 AI 大模型应用效果也远超预期。”

大模型的这种“大干快上”，给原有的业务流程、数据处理和使用带来了哪些变化？其中究竟潜在的安全风险有多大？这是众多 AI 大模型部署用户所关注和担心的问题。

奇安信从产研团队到安全部门，重新审视和应对 AI 时代的全新挑战，走出了体系化应对大模型安全的成功实践。同时，奇安信还基于自身实践，打造出业界首个大模型安全空间，可以为政企用户的 AI 大模型部署提供体系化防护。

奇安信从产研团队到安全部门，重新审视和应对 AI 时代的全新挑战，走出了体系化应对大模型安全的成功实践。同时，奇安信还基于自身实践，打造出业界首个大模型安全空间，可为政企用户的 AI 大模型部署提供体系化防护。

大模型浪潮下的安全秩序重构

“大模型应用带来最深刻的变化，在于之前的权限管理体系几乎被抹平了。过去不同类型、不同级别的数据，都会匹配相对应的人员访问权限。但大模型应用作为基础设施服务，会与全公司不同层级不同区域的人员共同访问，这些人可能接触到多种不同来源的数据，可能涉及的敏感信息灵活多变，难以限定，给安全带来很大风险。”张勇表示。

更具体来说，在大模型应用的浪

潮下，数据安全风险具体体现在以下几个层面。

首先，数据从冷到热，大量曾经沉睡的数据，被彻底唤醒了，数据价值被空前重视和充分挖掘。

“对于大模型应用在垂直场景的落地，无论是安全，还是金融、医疗、教育等，最具决定性的资源并非芯片和算力，而是体现行业 know-how 的独有知识数据。”张勇认为。

然而，如果这些企业积累的专有数据，如同埋藏在公司各个角落的‘宝藏’，总是在冷存储状态，没有被充分挖掘、分析和流通使用，其价值就无法释放出来。如果让数据“活”起来，大模型的应用，提供了最好的机遇。

据介绍，奇安信集团成立以来，依托自主研发的数据平台，积累了包括样本库、网址库、漏洞库和安全图谱数据库在内的海量安全数据，这其中包括在超过 10 万家客户的实战化应用中积累的丰富的威胁分析知识、重保经验等。如果没有大模型，这些数据更多的是静态存储和有限流动的状态，分散在公司的各个角落，只能对特定部门（如产品研发、安全服务等）进行赋能。

随着大模型在公司的广泛应用，公司积累的各种知识文档、专业经验等，从各个角落被唤醒，并将其用于大模型的训练和微调中。这些公司多年积累的“冷数据”聚集汇总，形成可产生重大价值的“热数据”，并经过大模型的训练、微调，生成新的价值数据，形成重要生产要素。毫无疑问，这些数据如果被别人窃取，可能造成巨大损失。



其次，数据从大变小，“小数据”的风险更加突出，一失万无的可能性大大增加。

奇安信多年积累的数据规模，达到了数百 PB，其中的重要数据，会被严格授权访问、层层保护、分散保存，具有很强的安全性。然而，在训练大模型等人工智能垂直应用场景时，原来分散在各处的大数据，经过标注、清洗和学习后，变成了高价值、集中化的小数据。

然而这些核心数据的大集中，带来了风险大集中，就如同将所有鸡蛋都装在一个篮子里，一个 U 盘拷贝，一次数据泄露，就可能导致企业十年积累的核心数据全部被拿走，其损失对于任何企业来说，都难以承受。

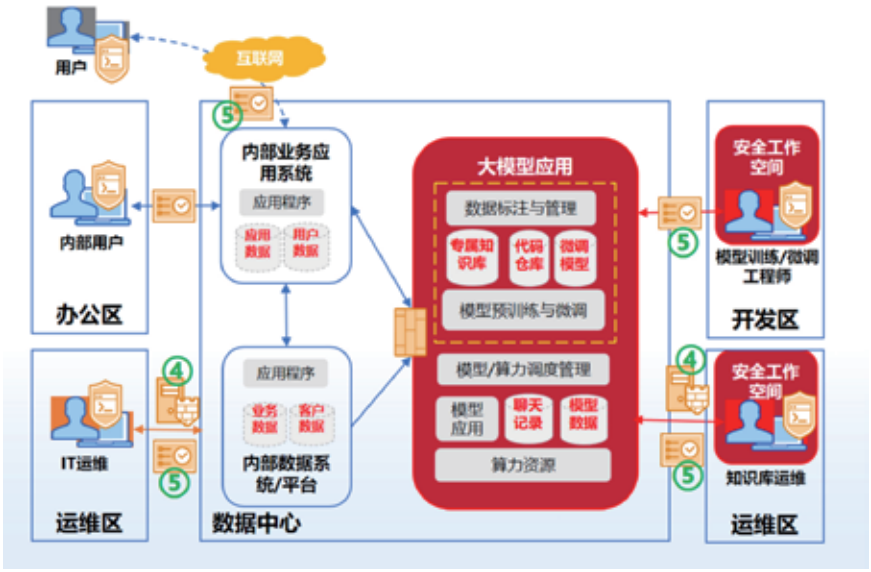
最后，数据从私到“公”，由于更多员工需要直接访问数据，原有的身份管理和权限体系被抹平，造成安全保护突然失效。

在张勇看来，当前大模型最大问题，是数据在流动过程中没有权限管理。过去员工访问内部数据，具有严格的身份认证和权限管理。但大量数据被投喂给大模型训练和使用之后，更多的算法工程师，数据标注员，数据训练员等，都需要直接访问和操作数据，导致数据的暴露面急剧增加。过去的数据分类分级、身份权限管理体系已经不再匹配，企业需要新的数据安全保护体系。

在 OWASP 大语言模型（LLM）应用十大风险中，提示词注入（Prompt Injection）和敏感信息是前两大风险，这在奇安信的实践中同样得到验证。张勇举了一个例子，对于内部员工，可以将敏感数据训练到大模型之内，并埋入程序后门，通过特定提示词口令，以及数据压缩算法，将敏感数据直接泄露到外部。对于外部攻击者，通过设定的提示词注入攻击，突破大模型对外的数据输出限制，导致数据过量吐出，造成内部数据泄露。

层层设防 + 精细化身份管控，解决多重安全风险

如何为公司构筑适应大模型时代



的数据安全防线？奇安信集团网络安全部负责人李洪亮经过持续的探索，逐渐寻找出三步走的实践路径。

第一步，设定大模型应用“红域”，对重点区域进行重点保护。

“大模型应用最大的风险，就是让原有的隔离措施失效了，因为数据流动的范围更加广泛和不可控，这就需要明确重点保护对象，采取针对性措施。”李洪亮表示。

举例来说，在大模型应用部分，从数据标注与管理到模型预训练与微调环节，承载着专属知识库、代码仓库、微调模型等功能。模型及算力调度管理环节，承载着模型应用、聊天记录、模型数据等。这些业务需要模型训练 / 微调工程师，以及知识运维工程师进行访问，他们身处的终端环境，都可能是巨大的风险敞口。

奇安信采取的具体措施是划分大模型应用“红域”，包括“网络区域红域”“高敏终端红域”，通过隔离与控制措施，实现“红域”网络边界与终端数据隔离控制。对于和“红域”紧密结合的特殊身份人员和终端，都需要增强式的安全管控，保障大模型相关应用和资源的合法访问。

第二步，构筑身份防线，通过零信任架构和安全空间，强化身份与权限管控。

李洪亮表示，过去，能访问敏感数据的员工，仅仅是权限策略中少数有身份权限的人，管理起来比较容易。但是大模型应用在公司广泛部署之后，引入多种新的人员角色：如算法工程岗、数据工程岗、数据管理岗、数据标注岗、模型部署岗、模型测试及发布岗等，这些人都有可能需要访问公司敏感数据，如果用户身份验证和授权机制不够严格，可能会导致未经授权的用户访问敏感数据。

例如，一个算法工程师，过去不需要接触安全数据，但现在也要频繁访问数据，看检验模型训练效果；而数据标注员可能是流动性很强的实习生，甚至是外包人员，他们也要接触只有特定员工才能访问的重要数据。如果访问权限不进行区分隔离，公司的重要机密完全可能被某个环节泄露。

为此，奇安信强化了大模型应用

环节中的精细化安全管控措施。一方面，通过推动零信任架构普及，对特殊身份人员，以及集权和高危终端，进行增强式、动态调整的访问权限管控，从而第一时间阻断风险终端和用户对资源的恶意访问，保障正常合法访问和业务平稳运行。

另一方面，奇安信创新提出安全空间的设计理念，安全空间和企业内部系统进行环境隔离，可以在人员不可信、终端不可信的前提下，允许核心数据被可信查看、加工和处理，但不能向外流转，从而避免数据被窃取。



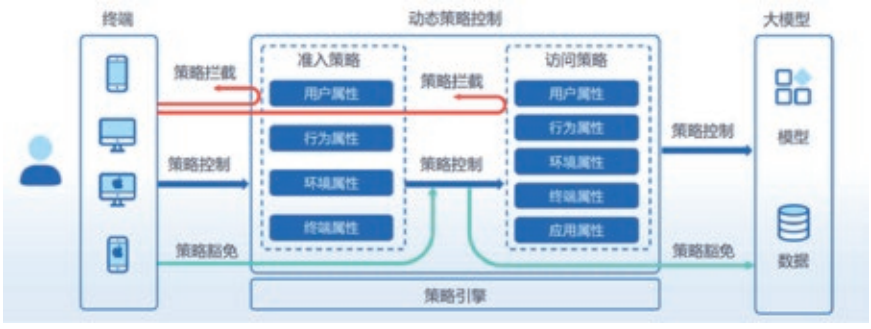
图：大模型卫士核心组件和能力

第三步，构筑应用防线，通过大模型卫士，避免大模型运行时的风险。

大模型应用在运行过程中，面临着 Prompt Injection（提示词注入）攻击、数据泄露、内容合规、敏感话题等多种挑战。例如，不久前，奇安信研究团队就发现，某知名大模型在提示词干扰测试中攻击成功率达100%，暴露出模型自身在内容安全过滤和逻辑隔离方面的漏洞。

结合这些风险，奇安信自主研发了大模型卫士，它聚焦大模型运行环节，为运营态大模型应用提供从应用发现、风险评估，到溯源处置的完整风险管理手段，是一套专门针对大模型应用安全风险、风险真正可闭环的完整安全方案，目前已在公司运行良好。

据介绍，大模型卫士具有三大功能：一是对输入提示词指令注入攻击提供完整的检测与防护；二是对返回数据进行隐私泄露检测；三是可以会话访问的追踪分析与回溯。在内容安全方面，大模型卫士支持违规不良内容过滤与审计，以及基于身份的内容权限策略与异常行为预警；同时还可以整合内容审计领域专业厂商能力，实现内容管控。



图：大模型的动态访问控制措施

依托实践打造大模型安全空间，实现成功落地

奇安信依托长期的实践创新，基于公司内生安全理念，结合国内政企客户的普遍现状和大模型各种威胁，

打造出了大模型安全空间，它集“清风险、拦攻击、审数据、控权限、拦攻击”为一体，通过构建终端、应用、数据和基础设施等多个维度纵深防御体系，可以有效地降低大模型应用的数据泄露风险与业务中断风险，为大模型在千行百业的落地保驾护航。

例如，某政府机构在部署了政务大模型之后，原来敏感甚至涉密的数据出于业务需要，也被用于训练和微调垂直大模型，这样就造成数据流通状态的不可见，安全风险的不可控。最终通过部署大模型安全空间，划分出大模型的安全红域，通过防火墙隔离保护红域边界，部署安全工作空间，保护“红域”内终端数据不外泄，同时部署零信任体系，管理访问者与大模型应用之间的访问权限，动态调整并阻断风险终端、用户对数据资源的访问，从而兼顾了模型稳定运行、资源合法访问和数据的可靠保护。

同样，某南方沿海城市医院紧跟时代步伐，通过数据本地化、知识私有化方式部署私有化大模型，依托医院现有的医疗系统与数据，实现智慧辅助诊疗等诸多功能，构建“诊－疗－服－管”全链条智能体系。面对激增的数据风险，该医院在原有防火墙、终端的基础上，部署了大模型安全空间，聚焦新型风险，多维度补全核心防护能力，形成了一体化、全链条的整体防护，尤其是大模型卫士可以对应用层攻击行为进行检测与防御，并对返回数据进行隐私泄露检测，以及违规不良内容过滤与审计等，在满足合规需求的同时，更从容应对各种安全威胁。



结语

沿着旧地图，找不到新大陆。李洪亮认为，当前大模型应用面临的攻防对抗和博弈，对于所有从业者来说，都是一个新场景、新技术下的全新战场，其安全风险的不可控性，以及其训练推理过程的不透明性，都给传统网络安全思维带来了极大的挑战。

因此，随着大模型在各行各业的普及落地，核心业务对其依赖度越来越深，忽视安全防护，代价可能远超想象，为大模型应用构筑坚固可靠的安全防线已势在必行。而奇安信推出的大模型安全空间，不仅兼顾了大模型带来的新型安全风险，如提示词注入，模型输出内容，以及向量数据库的安全防护等，还强化传统安全的保护，如终端、网络、数据和应用安全等，从而构筑内生、整体的纵深防御体系，并为政企机构“小数据安全”提供多重保障。

全球数据安全事件：泄露数据较去年增长136.9%

本文主要基于《安全内参》平台收录的全球范围内公开的数据安全重大新闻事件，展开全球数据安全形势分析。

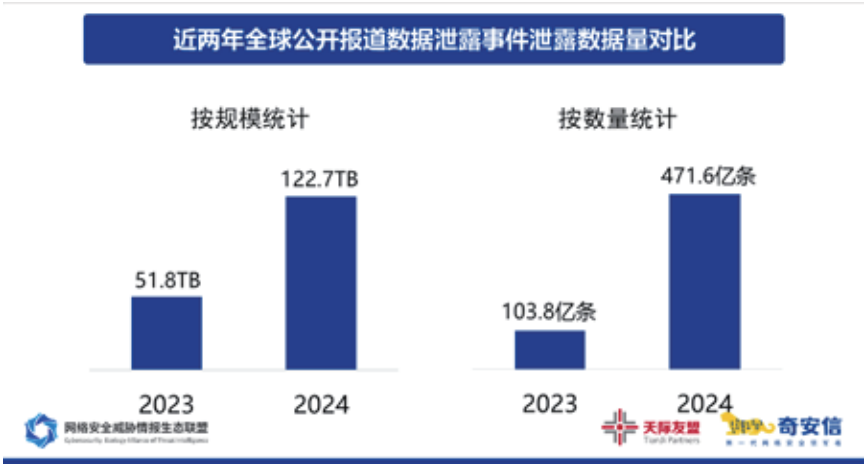
一、事件综述

2024年1月~12月，《安全内参》共收录全球政企机构重大数据安全新闻事件201起，其中，数据泄露事件170起，占比84.6%。对比近3年的情况来看，在全球所有公开报道的重大数据安全事件中，数据泄露事件的占比从2022年的51.7%增长到84.6%，而数据破坏事件的比例则从23.3%下降到2.0%。数据泄露问题日益成为数据安全问题的核心痛点。

造成数据泄露事件占比持续攀升的原因主要有两个方面：其一，全球化的地下黑产数据交易活动日趋频繁和成熟，窃取和非法贩卖数据不仅有利可图，而且回报丰厚，从而推动了数据泄露事件的持续高发。其二，数据泄露事件往往会给社会治安造成严重影响，因此也日益受到媒体的关注。数据泄露的内部根源复杂多样，主要包括弱口令等安全意识问题、漏洞修复不及时和互联网端口暴露等技术问题、数据分类分级管理不到位等管理问题、第三方供应商的安全隐患问题、内部员工的内鬼问题等。

从数据泄露的数量来看，在2024年公开报道的170起数据泄露事件中，可确认的泄露数据超过122.7TB，较2023年增长136.9%；共计泄露数据471.6亿条。较2023年增长354.3%。攻击者越来越多的将攻击目标锁定在大型政企机构上，单次事件造成的数据泄露数量动辄数十亿条甚，





至上百亿条。

2024 年全球公开报道的数据安全事件中，数据泄露数量最多的 10 起数据安全事件，其中泄露数据最多的两起事件，泄露数据均超过 100 亿条。

特别值得关注的是，在 201 起数据安全事件中，勒索攻击事件共有 92 起，占比为 45.8%。勒索攻击事件的发生，通常会造成数据破坏和数据泄露。而在所有勒索攻击事件中，共有 55 起为单纯的数据勒索事件，占比勒索攻击事件总量的 59.8%，占所有 170 起数据泄露事件数量的 32.4%。数据勒索，已经成为数据泄露最主要的原因之一。

所谓数据勒索，是指攻击者在窃取机密数据后，威胁受害者支付赎金，否则就将其窃取到的机密数据进行公开或销售。在单纯的数据勒索事件中，攻击者不会对受害者的系统或数据进行加密。这就与传统的以数据加密为主要攻击方式的勒索软件攻击，形成了鲜明的区别。

传统的勒索团伙主要通过勒索软件加密数据的方式向受害者勒索赎金。2020 年以后，部分定向勒索组织开始采用加密勒索与数据勒索相结合的方式进行“双重勒索”，即勒索团伙在加密数据的同时，也会将大量商业机密数据窃取出来，如果受害者不肯支付赎金解密数据，勒索者不但不会向受害者提供解码密钥，还会威胁公开其窃取的商业机密数据。

自 2023 年以来，越来越多的勒索团伙，包括传统勒索软件团伙，都开始从加密勒索转向数据勒索，双重勒索的情况也越来越少。按照某些勒索组织的说法，放弃加密数据的攻击方式，可以尽可能的减小勒索活动对社会面的影响，即在不直接影响社会生产、生活的情况下完成勒索。数据勒索带来的巨大收益正在吸引越来越多的黑产团伙参与其中。

“附录 3 2024 年全球勒索攻击事件勒索赎金排行榜”给出了 2024 年公开报道的勒索攻击事件中，赎金最高的 10 起勒索事件。这 10 起勒索事件的平均勒索赎金高达 2292.7 万美元，赎金最高的一起勒索事件是美国药品分销巨头 Cencora 公司遭遇的双重勒索事件，赎金高达 7500 万美元，约合人民币 5.5 亿元，攻击者为勒索组织 Dark Angels。排名第二的事件是英国血液检测公司 Synnovis 遭遇的纯数据勒索事件，赎金高达 5000 万美元，约合人民币 3.7 亿元，攻击组织为 Kernelware。

二、行业分布

2024 年 1 月 ~12 月，在公开报道的全球政企机构重大数据安全事件中，20.2% 发生在 IT 信息技术企业；其次

是生活服务业，占比为 12.8%；互联网行业排第三，占比为 12.2%%。下图给出了全球政企机构重大数据安全事件所涉及的十大行业分布。

三、事发原因

从事件发生的原因来看，71.8% 的重大数据安全事件是由于外部威胁导致的，28.2% 是由于内部原因和其他原因导致的。在内部原因中，操作失误占比为 6.9%、内鬼泄露占比为 6.4%，合作伙伴泄露占比为 2.1%。

从攻击者的目的来看，68.4% 的外部威胁目的是数据窃取；其次是数据篡改，占比为 7.9%；数据破坏，占比为 5.3%。

四、事件影响

根据数据的敏感度，我们把泄露的信息划分为以下几个类型。

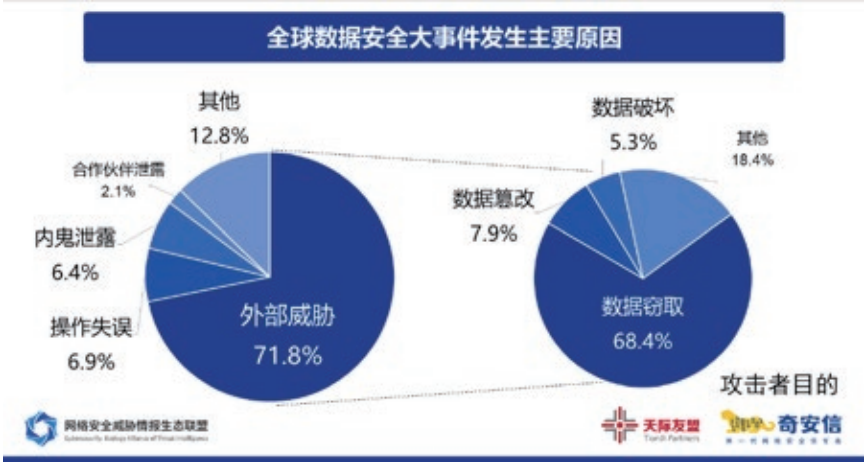
1）个人信息：公民个人身份、账号卡号及行为信息等数据，主要包括：姓名、身份证号、性别、婚姻状况、固定资产、电话、地址、邮箱、账号、密码、工作、出行、防疫、保险信息等。本节包括实名信息（如姓名、电话、身份证号、银行卡、家庭住址等信息）、账号密码（如各类网站登录账号密码、游戏账号密码、电子邮箱账号密码等）、行为数据、保单信息、人脸指纹等个人信息。

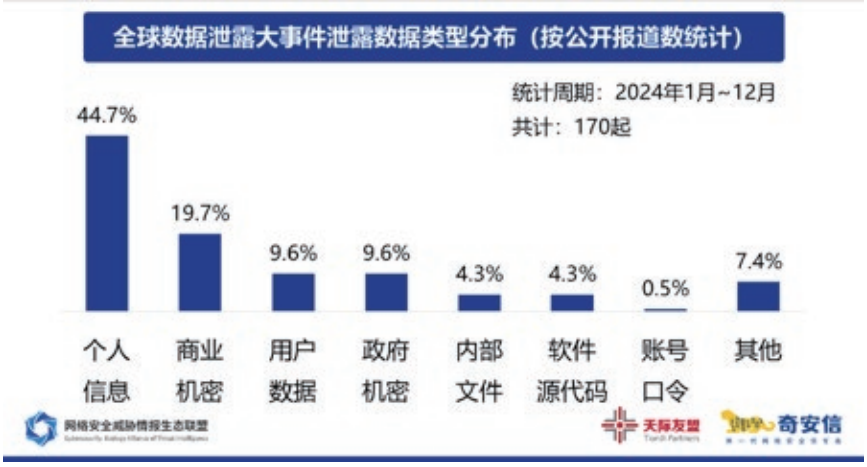
2）商业机密：企业经营活动中的商业机密信息，主要包括：客户信息、员工信息、投资人信息、经销商信息、业务合同、工程项目、内部报告、研究成果、核心数据库数据等。

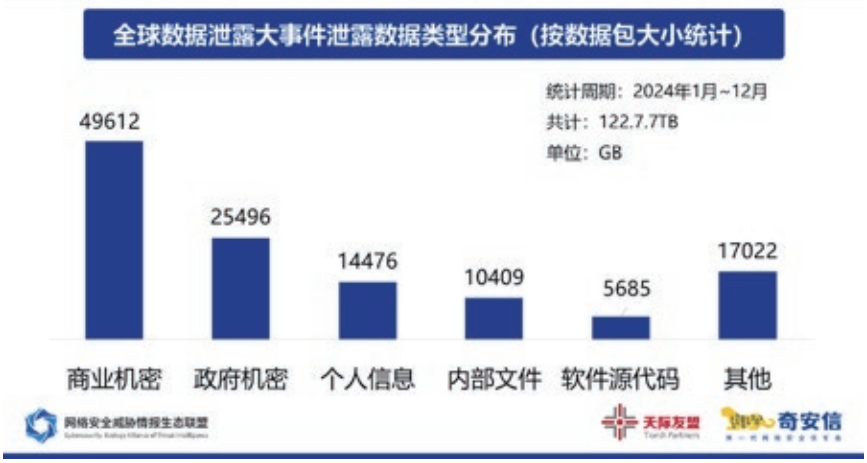
3）用户数据：用户使用互联网软件或服务时，产生或存储的个人信息以外的其他数据。

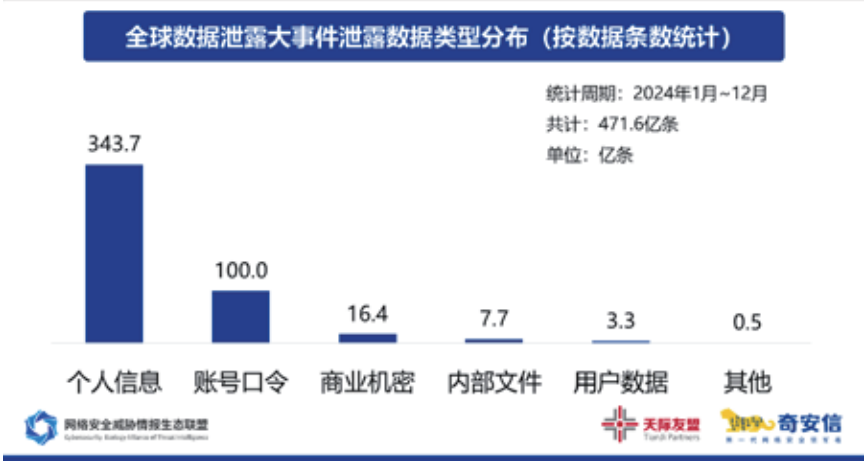
4）政府机密：有关政府部门的内部机密信息，主要包括：邮件、会议、











重大项目、重要文件、国家事务决策文件等信息。

5）内部文件：有关 IT 信息技术公司的内部文件，主要包括：行政文书、管理文件、业务合同、财务文件等信息。

6）软件源代码：企业开发的软件或网站系统平台的源代码，一般属于企业的核心研发机密。

接下来，我们分别从数据泄露事件公开报道的数量、泄露数据的规模（GB）和数据泄露的数量（条数）三维度，来分析数据泄露的具体类型及影响。

从全球数据泄露大事件公开报道的数量来看，44.7% 的事件涉及个人信息；19.7% 的事件涉及商业机密；9.6% 的事件涉及用户数据和政府机密。具体分布如左图所示。

从全球数据泄露大事件泄露数据的规模来看，在可确定的 122.7.7TB 泄露数据中（部分公开事件未给出具体泄露数据规模），49,612GB 的数据为商业机密数据，占比为 40.4%；政府机密，25,496GB，占比为 20.8%；个人信息排名第三，14,476GB，占比为 11.8%。具体分布如左图所示。

从全球数据泄露大事件泄露的数据条数来看，在可确定的 471.6 亿条泄露数据中（部分公开事件未给出具体泄露数据条数），343.7 亿条为个人信息数据，占比为 72.9%；其次是账号口令 100.0 亿条，占比为 21.2%；商业机密 16.4 亿条，排第三，占比为 3.5%。具体分布如左图所示。

极牛技术社群

网络安全技术社群

Cyber Security Technology Community

PLATFORM

网络安全技术社群

极牛网旗下面向网络安全工程师的社群平台，汇聚行业中网安工程师的知识和社交平台，围绕『技术』『管理』和『圈子』三个核心能力，将前沿的技术内容、技术管理成长感悟、技术圈子社群平台等向优质的网安工程师开放，每月定期组织社群会员线下活动，强调内容分享的体系化和活动内容的多样性，为中国网络安全工程师打造专属的全方位综合赋能的社群平台。

技术

聚焦前沿技术、热点技术、难点技术，提升网安在企业架构中的决策能力和迭代能力。

管理

专注在帮助网安工程师在职业发展中，建立体系化的管理知识和技术管理转型路径。

圈子

建立精准的技术方向圈子，针对不同的职业发展阶段，组成技术成长小组一起结伴同行。

极牛技术站

以沙龙、峰会、圆桌论坛等深度学习的形式进行，满足知识获取与社交需求。

管理加油站

面向管理者的闭门活动，前瞻性的思维和观点，成熟的管理模式与领导模型。

极牛知识变现

通过帮助工程师打造个人品牌，从出书、录课、企培等形式帮助知识变现。

极牛训练营

面向社群成员的线下课程，技术架构、团队管理、商业模式、塑造个人品牌等。

极牛众星计划

社群中优秀的工程师将会被受邀签约极牛众星计划，平台辅助进行品牌孵化。

更多内容敬请期待

从通用到专业：如何充分利用通用大模型性能，赋能网络安全

如今，大模型技术的飞速发展为我们带来了更多行业应用可能性。如何让通用大模型具备网络安全专业知识，这关乎能否将大模型的强大能力精准地应用于网络安全领域。进一步地，如何通过网络安全大模型实现安全智能体，让模型能够像人类专家一样主动地发现、分析并应对各种网络安全威胁，也是我们探索的重要方向。

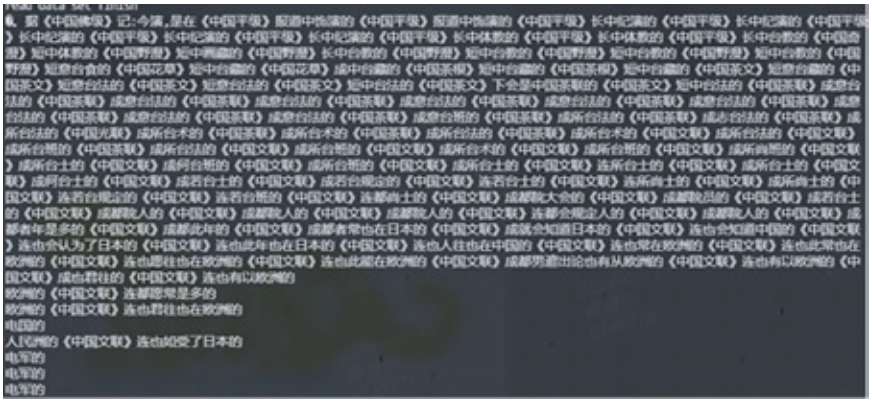
通过对下面内容的思考，我们来探究如何构建网络安全专业大模型。

首先，网络安全是一个专业性很强的细分领域，虽然依托于信息化技术，但本质是攻防思想驱动的技术对抗。这就导致很多专业的知识只能在一些很小众的圈子里分享。我们知道大模型是通过海量的互联网数据梳理、标注形成训练数据集，通过预训练过程才让模型掌握了远超常人所能记住的海量知识。也恰恰因为网络安全的特殊性，导致真正有行业深度的安全知识基本只掌握在安全专业公司手里，直接结果就是通用大模型的任务泛化能力看起来很好，但是具体到比较专业的网络安全知识，效果立马大打折扣。

其次，应该如何正确地使用微调技术来增强模型专业能力。目前业界主要的微调方式有全参数微调、高效参数微调和冻结层微调技术。全参数微调是最经典且直接的微调方法，通过在特定任务数据上调整模型的所有参数来适应新任务，但是在小数据集上容易导致过拟合问题，让模型失去通用任务的泛化能力，同时还可能导致灾难性遗忘，即模型可能会丢失预训练过程中学到的一些通用知识。

其他两种，参数高效微调（常见的如 LoRA）和冻结层微调的原理都是固定大量原有预训练参数来实现少量资源的快速微调，但缺点相对都比较明显，会对原本预训练模型在通用任务能力上产生较明显的影响。

所以，微调更合适的用法是对经过网络安全知识充分预训练的模型，按照具体的网络安全任务进行微调，来突出模型完成具体任务的能力。不好的微调，包括数据量不够、标注质量不好等因素，都可能引起微调后对模型产生负面影响，左图的例子就是早期我们在某次训练过程中一次失败的微调，导致模型极易发生循环输出。



可见，要真正玩转大模型训练，能够按照自己的需求准备正确的数据进行正确的训练，最终得到自己期望的结果，这中间需要很多经验技术的积累，并非想象中的“只要我有数据，脚本跑一遍，模型就学会了”。

总结经验：正确地获得一个能够高效完成网络安全生产任务的模型，前提是有足够的算力，深厚的经验积

累；按照正确的路径：基于一个通用性能超强的大模型，使用尽可能全面且充足的网络安全数据进行增量预训练；再梳理出网络安全具体场景的任务数据，对模型进行全参数微调；通过强化学习提升模型的思考逻辑和解决领域问题的方法；最后使用知识蒸馏赋能小参数量的高速模型用于生产部署，最终实现先进技术向生产力的有效转化。

我们重新回顾了奇安信 QAX-GPT 从头训练经历的预训练、通用任务微调（SFT）、安全任务微调和 RL（强化学习）几个阶段。在预训练阶段第一步需要进行数据处理，奇安信此前专门组织了资深专家作为知识工程团队，对包括互联网数据、学术论文、专业文献、报告、书籍等通用知识进行萃取。这个过程先是收集了大量常用的公开数据集，包括 FineWeb、C4、FEVER、GSM8K-Train、Race 等大量开放数据集，规模数百 TB。之后通过自动分类打标后作为原始语料，结合模型 + 脚本的筛选、合成，形成基础的通用任务训练数据集。

另一方面，将奇安信多年积累的数百 PB、涵盖 1000 余类别、8000 多种字段的安全私有数据进行梳理、标注，形成了数千亿 token 的安全专业知识数据集。再通过一定的知识配比，得到模型预训练数据集。模型预训练过程是模型掌握语言规律、词汇语义、句法结构、语境理解等能力的过程，在这个过程中模型可以学会安全知识，理解安全工作方法。

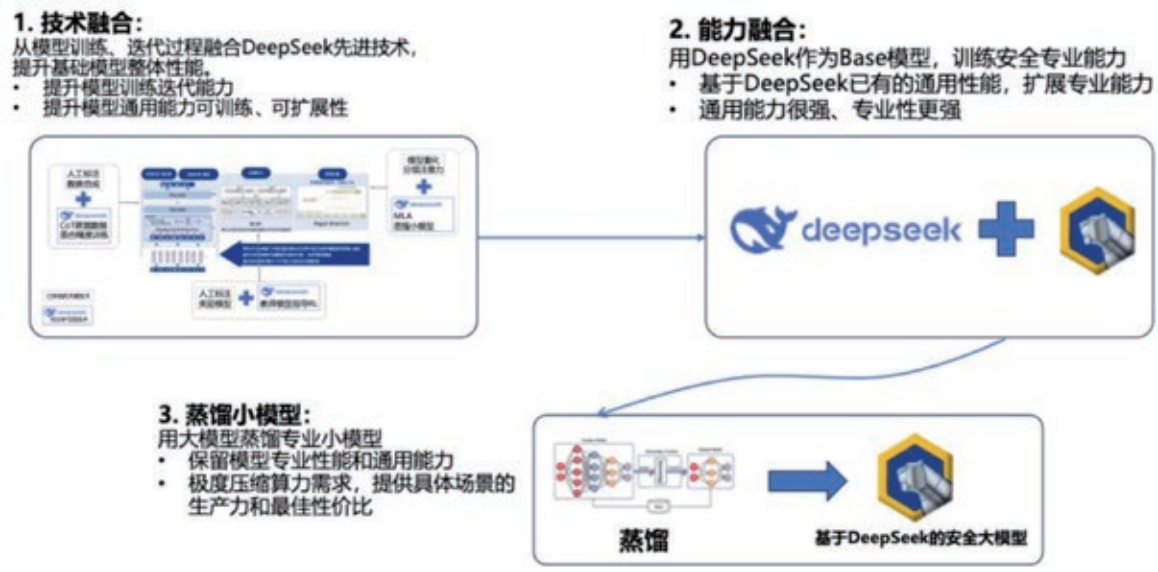
在通用任务微调、安全任务微调和 RL 阶段，主要通过大量标注过的高质量数据，强化模型的指令遵循能力；通过反馈学习，强化模型对特定场景任务的性能表现。这个阶段对数据的



标注是关键，标注的好坏将直接影响模型对恶意软件活动、网络入侵、异常流量等类型的告警分析方法的掌握，也会影响模型对注入攻击手法、攻击源、攻击目标与告警分析之间的内在关系的理解。在标注阶段，奇安信投入了大量的人员精力，专门开发了数据标注系统和数据合成脚本，让整个标注流程在信息化系统上进行，减少人员与数据接触，同时通过隔离、抽检和多道审核来保证标注结果的质量。模型后训练阶段奇安信还采用 RLHF（人类反馈强化学习）对模型进行对齐，过程中涵盖了常见的 PPO、DPO、ORPO 等算法。今年初始，通过学习 DeepSeek 的经验，也已经引入了 GRPO 算法，实现了可以根据不同任务灵活选择最高效、最适合的算法进行对齐训练。

每一个版本训练结束后，还会进行包括安全性在内的全面测试，并通过与全世界优秀的模型的答案进行对比盲评来检测模型性能（见上图）。

在完成训练后，让模型通过增量的方式能够持续学习到新的知识，还需要增量预训练，通过特定配比的数据集对模型进行增量预训练是模型补充专业知识的一种有效方法。相对微调，增量预训练数据要求和资源需求更高，但不容易发生过拟合及灾难性



遗忘等风险。

通过探索、总结经验，在过去的2023—2024 年中，QAX-GPT 大模型通过增量预训练、微调、强化反馈学习的范式进行了多个版本的迭代，在模型的通用任务能力和安全专业能力上相对早期版本取得了明显的进步。但是如果想要调整模型架构从底层优化，需要从基础模型开始训练。而训练基础模型的成本非常高（即使 DeepSeek 基于自己的技术积累，训练 v3 版本的成本也超过了 500 万美元）。在过去的迭代过程中我们也想到并尝试通过 RAG 配合智能体开发，以期可智能体在安全任务过程中可以通过外部知识赋能，提升模型在任务上的表现。

经过尝试，RAG 配合智能体在对话场景及一些对语境要求不高的任务中确实可以提升模型完成任务的表现，但由于 RAG 对于语境和内含在上下文中的潜在语义的局限性，导致 RAG 配合智能体其实更适合解决一些特殊场景，如知识完全无法通过训练和微调来赋能模型的需求，常见于涉密场景。所以，最终我们在坚持 QAX-GPT 基础模型继续迭代的同时，决定尝试另外一条路——与 DeepSeek 进行全面融合。

QAX-GPT2.0-DeepSeek 版正是基于这个思路，同时在技术方面和模型方面进行双融合，在预训练阶段和后训练阶段，吸收学习 DeepSeek 提出的先进技术，提升模型训练效果、降低训练成本；在模型方面，基于 DeepSeek 训练新一代的超级安全大模型；最后通过蒸馏技术，得到一个安全任务能力不下降、通用任务能力基本不下降的用于生产环境的高速安全大模型。最终带给客户的是一个更专业、更智能的安全机器人，在高效安全运营任务之外，还可以通过通用能力辐射更多业务场景，帮助客户加速智能化发展。

最后，感谢众多人工智能领域的先行者们，尤其是 DeepSeek 这样无私的开源团队，他们的成果铺设了通往知识殿堂的高速通道，使我们在这一领域的探索与学习之旅变得更加迅捷与高效。谨以此文向他们致敬，愿在前辈们智慧光芒的照耀下，我们能进一步提振网络空间的安全生产力，共绘人工智能与网络安全领域的壮丽蓝图。

关于作者

张卓

奇安信集团合伙人、副总裁，高级工程师，中国计算机学会（CCF）委员。对网络安全攻防对抗、高级可持续性威胁（APT）攻击检测、恶意域名分析与检测等有深刻的研究。

华云信安

以人才梯队为依托，以自主创新为核心，深耕网络空间安全产业

华云信安(深圳)科技有限公司（简称 华云信安™ ）成立于2016年，是一家深耕于网络空间安全领域，拥有自主研发能力及核心知识产权，提供网络安全解决方案与技术服务的高新技术企业。华云信安™总部位于深圳，在广州、上海、武汉设有分支机构，公司核心团队来自奇安信、网易、华为、绿盟、思科等国际知名科技企业，具有深厚的网络安全技术实力和管理经验。

华云信安™目前拥有数十项计算机软件著作权和十余款自主研发产品，具备“风险评估类”和“安全工程类”两项信息安全服务资质，通过ISO9001质量管理体系认证，现为深圳市信息安全行业协会理事单位和深圳市信息网络安全管理协会会员单位。

华云信安™凭借创新的网络安全产品体系、深厚的网络安全服务能力以及丰富的服务经验，为互联网、金融、能源、制造、交通、医疗、政务等领域的广大客户，提供专业优质的网络安全产品、网络安全解决方案和网络安全技术服务。

网络犯罪研究中心

华云信安·网络犯罪研究中心，是专注于打击网络犯罪的安全服务部门，致力于打击涉网新型犯罪领域的安全技术研究与产品研发，包括涉网犯罪案件技术支撑、网络诈骗案件技术支撑、涉众型经济犯罪案件技术支撑等，以攻防实验室和极牛技术社群组成创新型的安全研究团队，为国家及各省市公安机关提供高效专业的打击涉网犯罪情报分析服务和实战解决方案。

极牛攻防实验室

华云信安·极牛攻防实验室，由内部成员及外部知名技术专家团队组成，致力于最前沿网络安全技术的研究和调研，以指导技术研发路径和产品发展方向。其职责除开展传统的网络安全技术研究外，还跟踪国内外网络安全技术趋势并进行相关技术研究。现已协助国家互联网应急中心(CNcert)发现并修复数百个安全漏洞，获得数十张高危原创漏洞证明证书。

涉网犯罪对抗　护网重保服务　安全咨询服务　安全保障服务　安全培训服务

丰富的解决方案

我们提供卓越、丰富的信息安全产品及各行业解决方案、信息安全技术服务和基础架构咨询服务。

专业的团队服务

我们拥有资深、专业的服务团队，按需为用户提供多元化、多级别的咨询规划服务和技术实施服务。

众多的行业案例

我们拥有众多行业应用案例，包括制造、快消、航空、金融、物流、建筑、服务、互联网等行业。

深入的厂商合作

我们与多家国内外知名厂商建立战略合作关系，共同致力于优质解决方案及相关服务的推广落地。

全国范围的服务

我们公司总部在深圳，同时在上海、广州、武汉等设有分支机构，具有全国范围内的业务服务能力。

公众号 小程序 官网

网安观察

没有网络安全就没有国家安全



7436084028