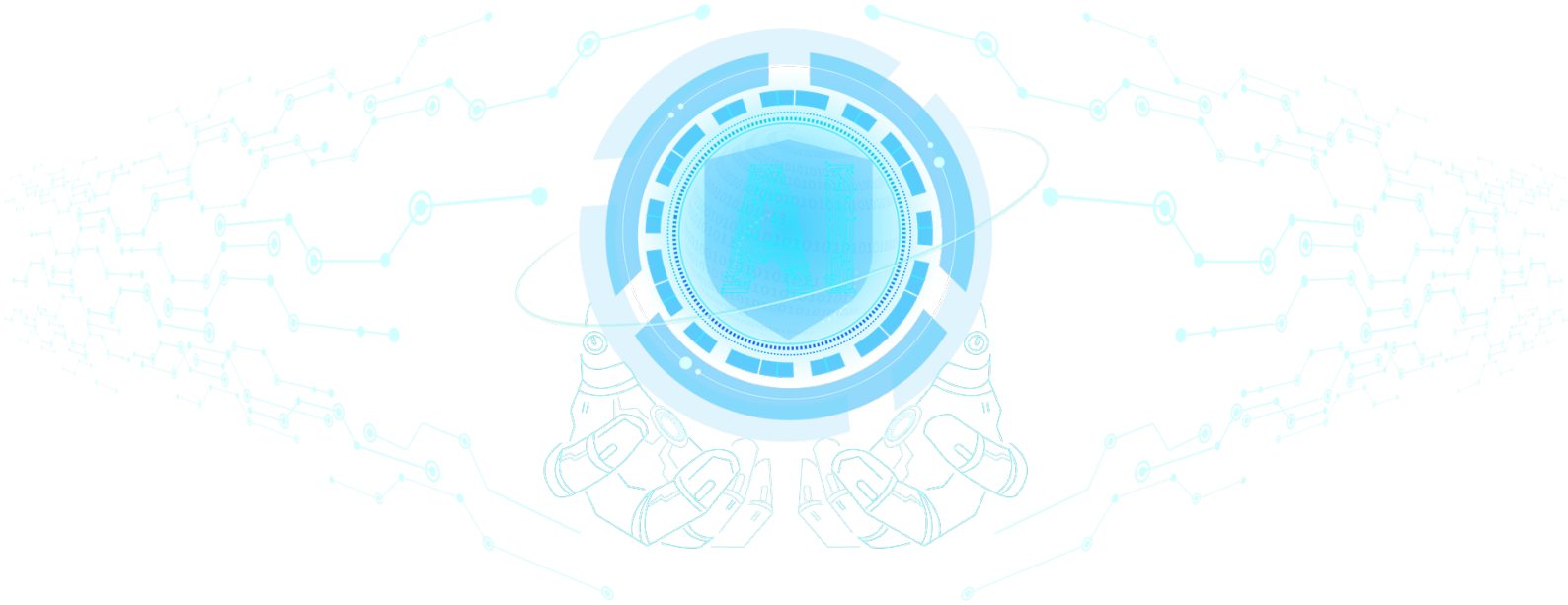
安全大模型技术与市场研究报告



数说安全
CYBERSECURITY REVIEWS

2024 年 6 月 25 日

目录

法律声明

本报告中的调研数据均采用行业公开信息、深度访谈、实地调研、桌面研究得到。本公司不承担因使用本报告而产生的任何法律责任。

前言

网络安全领域有一个“锤子”和“钉子”的理论：把网络安全中待解决的问题比喻作“钉子”，把解决问题的手段比喻作“锤子”。已经遇到过的问题叫“老钉子”，新遇到的问题叫“新钉子”，已知的解决问题的方法叫“老锤子”，新发明的方法叫“新锤子”。每当找到一把“新锤子”，人们会拿这把“新锤子”把“老钉子”都再砸一遍，看是否更好用，遇到“新钉子”也会先拿“老锤子”来砸一通，看是不是还管用。

生成式人工智能（AIGC）大语言模型（LLM，简称大模型）技术无疑是网络安全从业者拿到的一把“新锤子”，这把“锤子”已经在自然语言对话、写作、文生图、文生视频领域取得了令人震惊的效果，网络安全从业者自然是不会放过这样一个机会的，深信服、360、奇安信、安恒、绿盟、天融信、永信至诚、启明星辰等网络安全上市公司纷纷宣布自己发布了安全大模型产品，金睛云华等过去基于深度学习技术开发网络安全产品的公司也及时转向了 AIGC 大模型技术路线，新的安全创业公司也在纷纷采用人工智能大模型提升自身产品能力，如云起无垠在 Fuzzing 模糊测试方面，灵云数科在邮件安全方面都在使用人工智能大模型。

自 2010 年以来，我们拿到的“新锤子”包括机器学习技术、大数据技术、知识图谱技术等，用来砸网络安全的“老钉子”和“新钉子”，也取得了一定的成果，但网络安全的基本盘并没有改变：

- 攻防速度不对等：攻击者突破防线、偷走数据的速度远远快于防守方发现攻击、阻断攻击的速度，防守方的响应速度不够快；

  据：Unit 42 Cloud Threat Report - Volume 7, 2023, Unit 42 Engagement Experience 中披露的数据，目前的响应与处置时间已经提高到 6 天，比数年前的数

月有了很大的进步，但攻击者进攻得手的时间已经提高到数小时级别，攻防依然在速度上不对等。

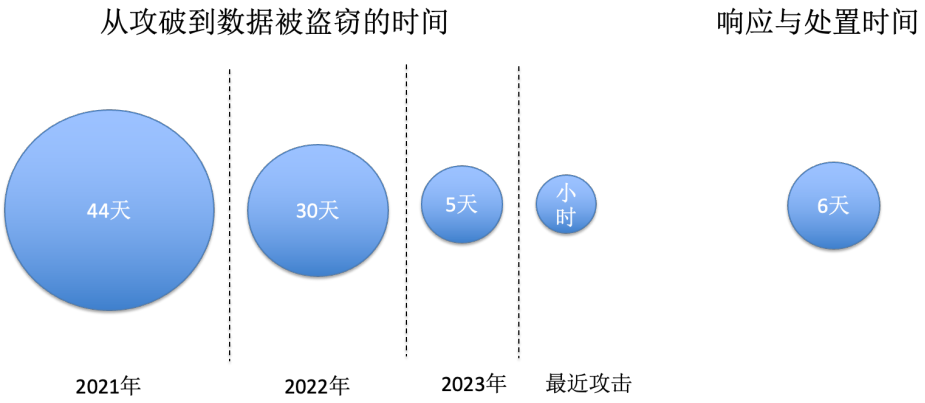


图 1 攻破到数据被盗窃的时间 VS 响应与处置时间

来源: Unit 42 Cloud Threat Report - Volume 7, 2023, Unit 42 Engagement Experience

- 现有网络安全防御方法的效率性价比不够高：现有的防御方式下，需要大量的安全工程师，中国、美国所公布的网络安全防御人才缺口都是几百万人，暂且不说培养出这么多网络安全从业者相当困难，雇佣这些人的成本也会是企业难以承受之重。
- 检出率不够高与误报率太高：基于规则的引擎对新型攻击的检出率不够高，规则引擎与 AI 引擎的误报率都太高。

AIGC 的异军突起，给了大家一个新的希望。本报告通过对基于 LLM 的 AIGC 在网络安全中的应用做了分析，希望给网络安全从业者提供一些有用的信息。

一、概述

自 2022 年底开始，以 LLM（大语言模型，简称大模型）为核心的 AIGC（生成式人工智能）带来了一场人工智能驱动的技术与产业革命，人工智能被广泛认为是可以改变“游戏规则”的战略性科技。国内开启了“百模大战”，国资委在 2024 年 2 月 27 日召开国有企业改革深化提升行动 2024 年第 1 次专题推进会上再次点题人工智能，要求国企结合自身技术资源禀赋和产业基础，科学决策纳入发展规划，加大投入，把资源用在刀刃上，有力提高国有企业战略支撑作用。

大模型也被寄予厚望，解决了多年来困扰网络安全行业的攻防不对等、安全专业人员严重不足的问题。

网络安全行业一直存在攻防不对等的问题：攻击者在暗处，防守者在明处；攻击者可以在任何时候发起攻击，防守者则需要 7X24 设防；攻击者 100 次攻击有 1 次成功即宣告成功，而防守者 100 次防守，1 次失守就算失败；全社会数字化转型背景下，需要防守的目标众多，而我国教育系统每年培养出的网络安全人才只有不到 3 万人，加上自学成才的人员，也还是远远难以满足构建网络安全防线的需求。并且由于人的反应速度问题、工作效率以及工作任务排队的问题，对安全告警的处理时间往往要 30 分钟以上，而攻击者在 30 分钟之内已经得手。

美国以微软为代表的企业在大模型爆发之初，就开始将大模型用于构建网络安全产品和服务，微软 2023 年 3 月 28 日即推出 Microsoft Security Copilot，Palo Alto Networks、CrowdStrike、Fortinet 等公司也快速跟进。国内深信服、奇安信、360、安恒信息、绿盟、天融信、永信至诚、金睛云华等公司也纷纷宣布自己已经推出，或即将推出基于大模型的

网络安全产品，在 2023 年年底，已经可以看到一些大模型赋能的网络安全产品的成功应用。

（一）主要发现

1. 供给侧视角：

- 在 AI 赋能网络安全概念下，各家所采用的 AI 技术五花八门，既有最新的 AIGC 大语言模型，也有深度学习技术甚至是机器学习技术，或者是多种技术的混用。
- 愿意提供纯软件部署方式，以解决在对华限售智能算力平台的大形势下，安全公司自身采购智算硬件的难题。
- 安全大模型的主要应用场景有：安全运营辅助、数据安全、威胁检测、电子邮件安全、开发安全、安全策略管理、渗透测试、安全培训。
- 安全大模型在各应用场景下的效果差异较大，在数据分类分级场景下取得的效果最好，能有几十倍的工作效率提升；在安全运营场景下，效果差异较大，与供应商安全大模型训练数据集与用户使用场景的匹配程度、供应商的技术水平都有关系。
- 产品有安全智能体与聊天机器人两种主要形态，目前聊天机器人是主要产品形态。
- 人工智能大模型技术会带来网络安全产品格局的大变化，AI 将会促成一系列网络安全产品的功能、性能提升，从而造成安全产品此消彼长的态势。

2. 需求侧视角：

- 中国用户与西方用户对产品形态上的要求差别很大，以美国为代表的西方世界市场普遍接受以 SaaS 服务为主要形态的服务，中国的客户则更愿意接受本地部署。

- 用户的主要诉求是通过提升自动化水平提升运营效率：缩短 MTTD、MTTR，希望能从天级，缩短到分钟级。
- 希望安全大模型结合威胁情报进行安全事件调查分析、取证。
- 用户希望通过安全大模型解决降低运营成本、提升运营效率问题。
- 用户希望通过安全大模型提升现有安全运营人员水平。
- 用户希望通过安全大模型能帮助总结安全事件。

3. 技术视角：

- 国内各厂商所采用的生成式 AI 技术，基本都是在商用或开源的基座大模型的基础上做预训练与精调。只有极个别厂商声称未来会考虑从头训练自己的安全大模型。
- 生成式 AI 与传统深度学习、机器学习技术相比，在威胁检测、加密流量分析上，检出率没有什么优势，但运算速度慢很多，大模型与小模型联合使用，会是现阶段比较好的选择。
- 在基座大模型的选型上有很高的趋同性，国产大模型选用通义千问的比例较高，国外的开源模型，LlaMa-2 和 MISTRAL 7B 模型的普遍评价比较高。
- 国产大模型算力依然在快速发展中，如何提供高性价比的训练、推理芯片，依然是现阶段所面临的挑战。

4. 市场视角：

- 政府部门对安全大模型十分关注，因地方政府建设了很多算力中心，算力问题也基本不成问题，是对安全大模型跟进速度最快的客户群之一。
- 金融机构对安全的大模型非常感兴趣，但在当前大环境之下协调 AI 算力资源有困难。

- 军方对安全大模型的本地化部署、训练、精调等方面有不同于政府和企业客户的要求。
- 安全大模型的价格战已经箭在弦上。

（二）推荐

1. 甲方单位网络安全与风险管理负责人应当

- 开始考虑使用安全大模型提升安全运营的效率，重新考虑人员规划与技术投入的比重。
- 根据自身 SOC/SIEM/XDR 等运营工具建设情况，选择相应的安全大模型产品或服务形态，如安全智能体或安全对话机器人。
- 开始考虑在数据分类分级、数据脱敏、数据防泄露中引入 AI 大模型用于提升数据安全管理自动化水平。
- 在邮件安全网关类产品选择上应优先考虑采用人工智能大模型的产品；
- 规划好算力问题如何解决。

2. 网络安全公司产品负责人应当

- 关注安全大模型对现有安全产品的颠覆性影响，这是网络安全行业的灰犀牛事件，将对全行业的产品构成产生深远的影响，有些产品可能会被安全大模型技术消灭，而有些产品的能力将得到大幅提升。
- 对安全大模型研发的技术难度有充分的准备，尤其是数据工程。
- 对训练数据的来源给予足够重视，确保合法合规。

- 时刻关注 AI 大模型底座的更新换代，及时切换到效果更好的 AI 大模型，并将各有特长的 AI 大模型用在最能发挥各自特长的地方。

二、 人工智能用于解决网络安全的历史

（一）深度学习技术出现之前，传统 AI 技术在网络安全领域中的应用

在深度学习技术成为主流之前，人工智能（AI）在网络安全领域的应用已经存在了很长时间。传统的 AI 技术在网络安全中的应用主要有：

1. 专家系统

专家系统在网络安全领域的应用历史可以追溯到 1980 年代末期至 1990 年代。专家系统是一种利用人类专家知识解决复杂问题的人工智能系统。在网络安全领域，专家系统被设计用来模拟安全专家的知识和决策过程，以便自动化地识别、防御和应对各种网络威胁和漏洞。

在早期，网络安全的专家系统主要集中在入侵检测系统（IDS）和病毒防护上。例如：

- 1987 年，安德森（Anderson）提出了一种基于规则的入侵检测系统的概念，这可以被认为是专家系统在网络安全中应用的早期例子之一。
- 1990 年代初期，首个商业化的入侵检测产品开始出现，它们采用了专家系统的技术，如 Sun Microsystems 的 Sun Screen SKIP 和 Digital Equipment Corporation 的 SecureWorks 等，这些系统能够利用已知的安全漏洞和攻击特征来识别潜在的安全威胁。

随着互联网的快速发展和网络攻击技术的不断演进，专家系统也面临着不断的挑战和需求，包括：

- 适应性和动态性：网络安全威胁持续变化，专家系统需要不断地更新规则和知识库以匹配新的攻击特征。
- 复杂性管理：随着网络环境变得日益复杂，专家系统需要处理更多的数据和情境，提高分析和判断的精确度。
- 集成与自动化：为了提高效率和效果，专家系统被要求更好地与其他网络安全工具和系统集成，实现自动化的安全防御。

2. 机器学习算法

2009 年前后，伴随机器学习技术的一波兴起，机器学习技术开始被应用于网络安全领域。

1）支持向量机（SVM）

支持向量机是一种监督学习的方法，用于分类和回归分析。在网络安全领域，SVM 可以用于恶意软件检测、网络入侵检测等场景，通过学习区分正常的数据和异常的数据。

360 公司的 QVM（Qihoo Vector Machine）引擎是全球第一款基于机器学习（SVM）技术的恶意软件检测引擎，2010 年 7 月投入实际使用，是 360 安全卫士的主引擎。未知恶意软件检出率>99%，误报率<0.01%

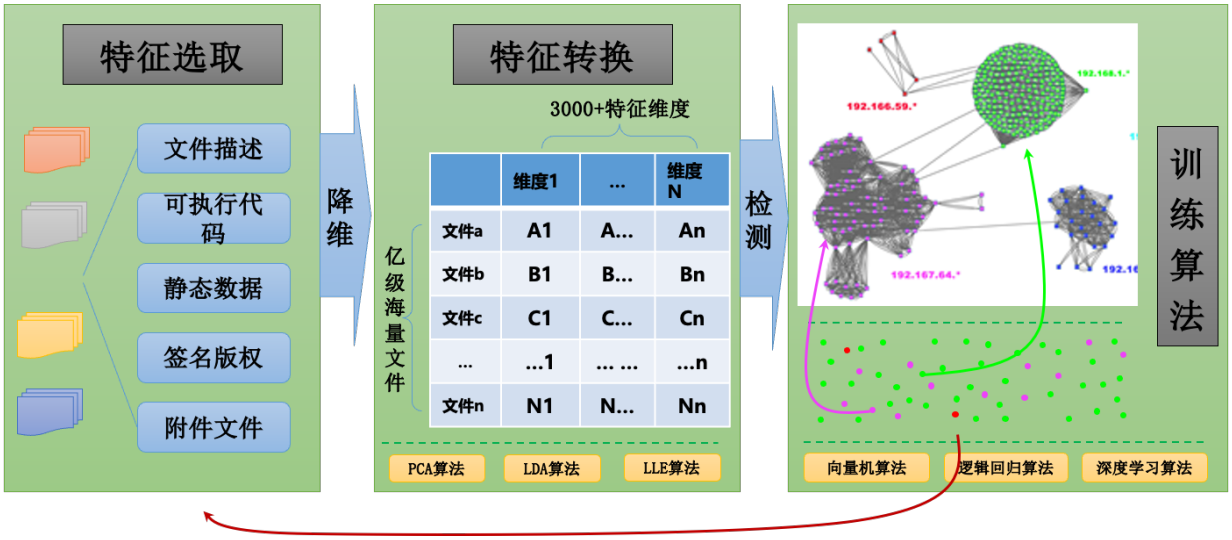


图 2 360 QVM 引擎原理

来源:360 数字安全

QVM 引擎的出现极大地提升了 360 产品在恶意软件检测领域的能力，尤其是对于那些新出现的、尚未有明确特征码的恶意程序有着良好的防御效果。2015 年 1~2 月，海外 10W+企业爆发加密敲诈者病毒(CTB-Locker/VirLocker)病毒家族，360 云安全中心，捕获 1700+变种，360QVM 在无任何人工干预和针对性训练情况下，识别率达到 99.3%

相对于传统特征引擎，因为其不具备明显可被绕过的特征，360 QVM 检出率衰减的速度要慢很多。

在后续的产品改进中，360 QVM 也采用了其它算法，包括试用过深度学习算法。

2) 决策树

决策树是一种用于分类和回归的方法，它通过从数据特征中学习决策规则来进行预测。在网络安全中，决策树可用于构建入侵检测系统，通过分析网络流量和系统日志来识别恶意行为。

3）遗传算法

遗传算法是受生物进化理论启发的搜索算法，适用于解决优化问题。在网络安全领域，遗传算法可以用于密码破解、安全策略优化等问题，通过模拟自然选择和遗传机制来逐渐逼近最优解。

4）模糊逻辑

模糊逻辑系统能够处理不精确或不确定的信息，对于分析含糊不清的网络安全数据特别有用。它可以应用于异常检测和威胁评估，通过模糊规则来识别和量化网络威胁的严重性。

5）贝叶斯网络

贝叶斯网络是一种表示变量间依赖关系的图形模型，可以用于概率推理。在网络安全中，贝叶斯网络能够帮助分析和预测网络攻击的可能性，通过计算不同安全事件之间的条件概率来评估风险。

总体而言，这些传统的 AI 技术主要集中在基于规则和模式匹配的方法，以及有限的机器学习算法上。它们对特定类型的网络安全威胁和攻击场景有一定的识别和防御能力，但在处理未知威胁、自适应攻击以及大规模复杂数据时，可能不如深度学习技术那样有效。随着深度学习技术的发展和应用，网络安全领域的防御能力和智能水平得到了显著提升。

（二）深度学习技术在网络安全领域的应用

深度学习技术在网络安全领域的应用日益增加，因为其强大的数据处理和模式识别能力使它成为增强网络安全措施的重要工具。以下是深度学习在网络安全领域的一些应用与研究成果：

1. 恶意软件检测与分类

通过使用深度学习模型，如卷积神经网络(CNN)和循环神经网络(RNN)，可以有效识别和分类各种恶意软件。这些模型可以从文件或程序的原始二进制数据中学习特征，以识别潜在的恶意行为。

2. 入侵检测系统

深度学习模型能够分析网络流量数据，以检测异常行为或入侵企图。通过学习正常网络行为的模式，它们可以识别出偏离这些模式的潜在攻击，包括零日攻击和先进的持续威胁(APT)。

3. 钓鱼网站识别

深度学习被用于自动识别钓鱼网站，通过分析网站的 URL、网页内容和其他元数据，深度学习模型可以检测出伪装成合法网站以窃取用户信息的钓鱼网站。

4. 域名生成算法(DGA)检测

网络犯罪者使用 DGA 来生成大量域名，以用于命令与控制服务器的通信。深度学习模型可以识别这些算法生成的域名的模式，帮助拦截与恶意软件的通信。

5. 基于行为的欺诈检测

在金融和电子商务领域，深度学习用于分析用户行为和交易模式，以检测和预防欺诈行为。这些模型能够识别出异常的购买模式和登录行为，从而防止欺诈发生。

许多学者和机构都在这个领域作出了重要贡献。例如：

- 2017 年，Saxe 和 Berlin 通过使用深度学习模型在恶意软件分类和检测方面取得了显著的进展，证明了深层神经网络在识别恶意文件方面的有效性。
- 在入侵检测领域，研究者使用深度学习技术开发了多种模型，如卷积神经网络(CNN)和长短期记忆网络(LSTM)，在多个标准数据集上取得了优异的检测性能。
- 针对钓鱼网站的检测，有研究者使用深度学习分析网站的特征和用户行为数据，有效提高了钓鱼网站识别的准确率。

这些应用和研究成果展示了深度学习技术在提高网络安全防护水平中的潜力和价值。随着技术的不断进步和更多数据的可用性，深度学习在网络安全领域的应用和研究成果将继续增长。

6. 数据安全

数据安全领域存在很多还没有别解决好的问题，比如数据分类分级，过去一直是小号大量人力的工作，并且难以做到伴随业务系统的升级做到及时与准确的数据分类分级，过去几年中业界尝试用深度学习技术辅助进行数据的分类分级并取得了一定的效果。

DLP 是数据安全中面临的另一个难题，过去所采用的数据指纹技术有识别准确率、误判率等一些列问题，采用深度学习技术进行内容的判定也是过去几年的努力方向之一。

(三) 知识图谱在网络安全领域的应用

知识图谱作为一种组织和利用知识的技术，其在网络安全领域的应用已经受到重视。具体的应用情况主要体现在以下几个方面：

1. 威胁情报分析

知识图谱能够整合来自不同源的威胁情报数据，如漏洞库、黑客工具、攻击行为等，通过构建实体及其关系，使安全分析师能够更有效地理解和分析威胁，从而更快发现潜在安全威胁和攻击手法。

2. 攻击检测和响应

利用知识图谱，可以构建网络攻击的模式和行为特征图谱，通过实时分析和匹配，提高异常行为和潜在攻击的检测速度和准确性。此外，通过知识图谱的推理机制，可以实现对攻击行为的预测和提前响应。

3. 安全态势感知

知识图谱可以整合和关联网络设备、应用、用户等多源安全数据，建立全面的网络安全态势图谱。通过对图谱的分析，帮助企业从宏观角度理解和评估其网络安全状况，发现系统薄弱环节，为制定安全策略提供决策支持。

4. 漏洞管理

通过构建涵盖漏洞信息、受影响产品、修复策略等信息的知识图谱，可以帮助安全团队有效管理漏洞信息，快速响应和处置已知漏洞，减少潜在风险。

5. 安全知识的教育与培训

知识图谱在整合和组织大量安全知识方面具有明显优势。利用知识图谱提供的结构化安全知识，可以为安全教育和培训提供丰富、直观的学习资源，提高教育和培训效率。

总的来说，知识图谱的应用提高了网络安全领域的数据分析效率和决策质量，为安全防护提供了更加科学、系统的支持。随着人工智能和大数据技术的不断发展，知识图谱在网络安全领域的应用将会越来越广泛和深入。

(四) AI 技术在网络安全领域的应用总结

1. 异常检测（Anomaly Detection）

AI 能够学习正常的网络行为模式，并通过实时监控，快速识别出偏离正常模式的活动，即异常行为，这些行为可能表明安全威胁，如入侵、恶意软件传播等。

2. 恶意软件和病毒检测（Malware and Virus Detection）

通过对恶意软件特征的学习，AI 能够识别出已知和未知的恶意代码。这种能力特别应对零日攻击（Zero-day attacks）十分关键，因为这种攻击利用的是之前未知的漏洞。

3. 垃圾邮件和钓鱼攻击过滤（Spam and Phishing Detection）

AI 可以用来分析电子邮件的内容、结构和发送模式，以识别和过滤垃圾邮件和钓鱼邮件。这有助于减少企业和个人受到的欺诈性信息攻击。

4. 身份认证和访问控制（Identity Authentication and Access Control）

人工智能可以提高身份验证过程的安全性，例如，使用生物识别技术（如面部识别、指纹识别等）进行身份验证，进而对用户进行更精确的访问控制。

5. 网络流量分析（Network Traffic Analysis）

AI 可以帮助分析大量的网络流量数据，以识别潜在的安全威胁，如分布式拒绝服务（DDoS）攻击、网络扫描等。对于加密流量分析，AI 是当今使用的主流分析手段。

6. 安全策略管理（Security Policy Management）

随着网络环境的不断变化，AI 可以帮助自动更新和维护安全策略，确保策略的及时性和适应性，减轻人工维护工作量。

7. 自动化响应（Automated Response）

在检测到威胁后，AI 可以帮助自动化响应流程，例如隔离受感染系统、阻断恶意通讯、甚至反向追踪攻击源等，提高响应速度和效率。

8. 欺诈检测（Fraud Detection）

在金融服务领域，AI 能够学习和识别欺诈交易模式，帮助机构预防信用卡欺诈、账户劫持等行为。

随着技术的发展，人工智能在网络安全领域的应用还将持续扩展，提供更加精准、高效的安全保障措施。

9. 数据安全（Data Security）

AI 辅助进行数据分类分级、DLP 数据防泄露中内容的检测。

（五）前大模型时代 AI 在解决网络安全问题上遇到的问题

1. 误报率

检出率与误报率是网络安全产品最重要的两个指标，而用户对误报率的容忍度比检出率更低，原因在于高误报率，会造成以下几种严重后果：

工作效率影响：高误报率意味着大量正常活动被错误地标记为恶意，这将导致安全团队花费大量时间和资源去调查并处理这些非威胁事件。这不仅分散了安全团队对真正威胁的关注，而且降低了工作效率。

系统性能影响：如果安全系统频繁地因误报而触发警报，可能会对网络或系统的性能造成影响。例如，对合法软件的误报可能导致该软件被限制或阻止运行，这会干扰正常工作流程，影响用户体验和生产力。

用户对安全检测系统信任度下降：频繁的误报会导致用户对安全检测系统的信任度降低。当用户习惯于看到大量的误报时，他们可能会开始忽略所有的安全警告，包括真正的威胁，这种现象被称为“警报疲劳”。

业务影响：在某些情况下，误报可能直接影响到业务运作。例如，如果一个安全检测引擎错误地将企业的关键业务软件标记为恶意并进行隔离或删除，这可能导致业务流程中断，造成经济损失。

因此，为了最大限度地提高网络安全防护体系的有效性和效率，同时最小化对业务运行的负面影响，对网络安全检测引擎的误报率提出严格要求是十分必要的。

2.数据质量和可用性

网络安全数据往往包含大量的噪声和不相关信息。而且，获取高质量、标注良好的训练数据往往非常困难和昂贵。比如使用机器学习技术训练的恶意代码检测引擎，需要的训练样本是数百万级别，并且要持续更新训练数据，以对抗最新型的攻击手法。

而对于特定类型的攻击（例如 APT 攻击）事件中，攻击样本相对于正常行为样本通常是非常稀少的，则很难选取到足够数量的训练样本。这种极度不平衡的数据分布给机器学习模型的训练带来了挑战，因为模型很容易偏向多数类（即正常行为），而忽视少数类（即攻击行为）。

3. 模型泛化能力

网络攻击的模式和策略在不断进化，这意味着即使是经过训练的模型也可能很快过时。模型需要不断更新以适应新的威胁模式，但这在实践中很难做到，导致模型的泛化能力受到限制。

4. 可解释性问题

尽管传统的机器学习方法能够在某些场景下有效检测到网络威胁，但这些模型通常被视为“黑箱”，难以理解其内部决策过程和逻辑。这在网络安全领域尤其成问题，因为安全分析师不仅需要知道一个行为是否恶意，还需要理解为什么是恶意的以便进行适当的响应。

5. 实时性能

网络安全系统需要实时或接近实时地监测和响应威胁。这对于许多传统 AI 模型来说是个挑战，因为它们可能需要显著的计算资源来处理大量数据，并且在检测到威胁时可能存在延迟。

6. 人工智能自身的安全问题

在机器学习模型中，对抗性攻击是指通过精心设计的输入来欺骗模型做出错误决策的技术。在网络安全领域，攻击者可以使用对抗性技术来规避检测。训练模型抵抗这类攻击是一项挑战，尽管近年来这一领域已取得了显著进展。

7. 人工智能人才稀缺

因为 AI 应用范围的迅速扩大，人工智能人才处于异常紧缺的状态，但人工智能人才的培养周期长，短时间之内问题难以缓解。

三、 大模型带来的 AI 驱动安全

大语言模型（例如 GPT-3.5、GPT-4、BERT、LlaMa2/LlaMa3、MISTRAL 等）在人工智能领域的发展中起到了革命性的作用，它们对人工智能技术的进步产生了深远影响。这些模型通过在大规模数据集上进行训练，能够理解、生成、翻译文本，甚至进行一定程度的推理。它们对人工智能技术的影响主要体现在以下几个方面：

（一） 大模型带来了哪些新可能性？

1. 自然语言处理能力的提升

大语言模型极大地提高了机器对人类语言的理解能力。之前，机器理解文本主要依赖于关键词或简单的语法规则，容易受限于语境的复杂性。大语言模型利用深度学习技术，能够理解语言的细微差别和语境，这对于复杂的自然语言处理任务（如情感分析、文本摘要、对话系统等）有重大意义。

2. 多种 AI 任务性能的提升

不只限于自然语言处理：大语言模型还能在跨语言翻译、文本生成、知识问答等多种任务上展现出优异的性能。它们通过大规模的预训练，掌握了丰富的世界知识和通用逻辑，这使得它们能够在没有特定域知识的情况下，也能生成可靠和自然的回答。

降低了人工智能应用的开发门槛：大语言模型的出现使得不具备深厚技术背景的开发者也能够轻松创建出高质量的人工智能应用。通过 API 调用或者轻微调整预训练模型，即可应用在各种场景中，极大地节省了从零开始训练模型的时间和资源成本。

3. 推理和逻辑

它们能够展示一定程度的推理能力，如因果关系推理、常识推理和解析较复杂的问题。例如，LLM 可以基于给定的信息进行分析，预测可能的结果或解释事件之间的关系。

上下文理解：LLM 能够根据上下文信息做出反应，理解对话历史中的细节，并据此调整其回应。

跨领域知识：这些模型通常接受广泛的训练数据，能够处理多种类型的问题和任务，跨越不同的知识领域。

4. AI 驱动的网络攻击

学术界和工业界在人工智能大语言模型辅助的网络攻击上已经取得了一些显著的进展。这些研究主要集中在如何利用大模型，如生成对抗网络(GANs)、Transformer models 等技术，以提高网络攻击的自动化程度、隐蔽性和效率。以下是几个研究领域的具体进展:

自动化攻击生成：通过训练模型理解和生成攻击代码的能力，研究者们已经能够自动生成针对特定系统或软件的攻击代码。例如，使用 Transformer 等自然语言处理技术，可以对已知的攻击策略和漏洞信息进行学习，进而生成专门针对新发现的漏洞的攻击代码。

钓鱼邮件与社交工程攻击：利用人工智能对大量数据的处理能力，攻击者可生成更加精准和难以区分的钓鱼邮件或社交工程攻击信息。例如，通过分析受害者的社交媒体行为，机器学习模型可以创建定制化的消息，显著增加欺骗成功率。

渗透测试的自动化：通过 AI 技术，可以自动化某些渗透测试过程，识别系统的弱点和漏洞。AI 模型可以被训练用以模拟攻击者的行为和策略，更高效地发现并利用系统漏洞。

逃避检测：通过使用机器学习模型，攻击者可以设计出能够绕过现有安全检测系统的恶意软件。例如，生成对抗网络(GANs)可以用来生成能够逃避恶意软件检测的变种。

对抗机器学习模型：针对特定的机器学习模型进行攻击，以破坏模型的正常工作或欺骗模型做出错误决策。这类研究不直接攻击网络系统，而是攻击系统内部用于安全防御的 AI 模型，例如通过对抗性样本来欺骗图像识别系统。

5. AI 驱动的风险识别

人工智能大模型在风险识别方面的研究也取得了很大进展：

数据分析与理解： AI 大模型可以处理和分析大量数据，并从中抽取有用信息，有助于识别潜在的风险点。比如在金融市场分析、网络安全威胁监测、保险欺诈检测等领域。

模式识别：这些模型通过深度学习技术，能够识别数据中的复杂模式和趋势，包括异常检测，这在预测金融市场的风险、监测网络异常行为等方面尤为重要。

自然语言处理（NLP）：AI 大模型在自然语言处理方面表现突出，能够理解和分析人类的语言。这意味着它们可以从新闻报道、社交媒体帖子、专业报告等中识别潜在的风险信息和信号。

动态学习与适应：这些模型能够从新的数据中学习，不断调整其分析和预测。这对于面对不断变化的风险环境（如网络安全威胁不断演变）尤其重要。

决策支持：AI 大模型还可用于提供基于数据的决策支持，通过高质量的风险分析报告，辅助专业人士和决策者做出更明智的选择。

预测建模：利用历史数据，AI 大模型可以构建预测模型，预测特定事件的风险级别。

在风险识别领域，AI 大模型利用其强大的数据处理能力、模式识别和自然语言处理等技术，能够帮助相关领域高效、准确地识别和评估各种风险，为决策提供科学依据。随着技术的不断进步和模型的不断优化，它们在风险管理方面的应用将更加广泛和深入

6. 新业态的出现

大语言模型的能力激发了新的商业模式和服务的出现，从内容创作到客户服务，再到个性化教育，各行各业都在探索如何利用这一技术创造价值。同时，为了应对这些模型带来的挑战，也催生了对数据隐私保护、算法伦理、模型可解释性等方面的技术创新。

（二） 产业界的热点方向

1. AI 赋能的威胁检测产品

人工智能在网络安全威胁检测产品的应用从多个维度提升了检测的效率和准确性，具体到恶意文件检测、攻击流量检测、用户和实体行为分析(UEBA)、以及加密流量分析这几个方面，其应用情况如下：

1）恶意代码检测

人工智能，尤其是深度学习和机器学习算法，在恶意代码检测中扮演着核心角色。这些技术能够分析文件的静态特征（如二进制代码结构）和动态行为（运行时的操作序列），学习历史上已知恶意软件的模式，并基于此识别新的、未见过的恶意文件。AI 能够快速适应恶意软件的变种，对于零日攻击的检测尤其有效。通过持续训练，模型能够不断提升对恶意文件的识别准确率，减少误报和漏报。

随着大模型的兴起，业界也在尝试采用大模型来进行恶意文件的检测，比如华清未央公司所发明的 MLM 大模型（机器语言大模型），就可以用来进行恶意代码的检测。

2）攻击流量检测

在攻击流量检测中，AI 和机器学习技术通过对网络流量的实时分析，能够识别出异常的数据包和通信模式，这些往往与网络攻击相关。AI 模型能够处理大量数据流，学习正常网络行为的复杂模式，并在此基础上识别出偏离常态的流量，比如 DDoS 攻击、恶意扫描、数据泄露尝试等。通过实时分析和模式匹配，AI 能够即时触发警报并采取防御措施，保护网络免受攻击。

业界现在有公司在开发 AI 智能体（AI Agent），用于做攻击流量的检测与自动处置。

3）用户和实体行为分析(UEBA)

UEBA 利用 AI 技术，尤其是无监督学习算法，来分析用户和系统实体的行为模式，以识别出异常行为。通过学习每个用户的历史活动，如登录时间、访问权限使用、数据访问量等，AI 能够建立正常行为基线。任何偏离这些基线的行为都可能被视为潜在的安全威胁，如内部威胁、账号劫持等。这种基于行为的分析方法能够在攻击发生之前预警，增强网络安全态势感知能力。

4）加密流量分析

随着加密技术的普及，加密流量成为网络攻击隐藏的一种手段。人工智能在此领域的应用旨在通过分析加密流量的元数据（如流量大小、时间模式、连接频率等），在不解密内

容的情况下识别异常流量。结合机器学习模型，系统可以学习加密流量的特征，识别出与恶意活动相关的模式，如恶意软件命令与控制(C&C)通信、数据泄露等。这种方法在保护用户隐私的同时，提高了对加密流量中隐藏威胁的检测能力。

综上所述，人工智能在这些领域的应用极大增强了网络安全威胁检测的智能化水平，使得安全系统能够更加精准地识别和响应各类威胁，降低了人工审查的负担，提升了整体网络安全防护的有效性。

2. AI 赋能网络安全运营

AI 大模型在网络安全运营中的产业实践，尤其是在告警降噪、攻击研判、自动响应与处置方面，展现出了显著的优势和潜力，具体表现在以下几个方面：

1）告警降噪

智能告警过滤与分类：传统的安全运营中心常常面临大量告警信息，其中许多是误报或低优先级事件。AI 大模型通过学习历史数据，能有效识别并过滤掉这些无关紧要的告警，仅将真正需要关注的高风险事件呈现给安全分析师，极大地减少了噪音，提高了响应的针对性。

上下文关联分析：大模型能够分析告警之间的关联性，结合时间序列、用户行为、网络流量等多种因素，为告警提供更丰富的上下文信息，有助于快速判断告警的真实性和严重程度。

2）攻击研判

复杂攻击模式识别：通过深度学习算法，AI 大模型能够识别出隐藏在大量数据中的复杂攻击模式，包括零日攻击、高级持续性威胁(APT)等难以用传统规则发现的威胁，提高了攻击研判的准确性。

自动化威胁狩猎：大模型驱动的威胁狩猎能力，可以在大规模数据中自动搜索潜在的恶意活动迹象，无需预设规则，而是通过模式识别自动发现异常，加速了威胁的发现过程。

3）自动响应与处置

自动化剧本执行：基于 AI 决策，大模型可以触发预设的安全响应剧本，自动执行隔离受感染设备、关闭特定端口、发送警告邮件等一系列响应措施，快速遏制安全事件的发展。

自适应安全策略调整：根据实时分析结果，大模型能够动态调整安全控制策略，比如加强特定区域的监测力度或临时限制某些高风险服务，实现更灵活的防御布局。

交互式辅助决策：在某些情况下，大模型还可以为安全团队提供决策支持，通过生成详细的事件分析报告，提出可能的处置建议，协助分析师作出更为精确的判断。

这些实践证明，AI 大模型已成为网络安全运营中不可或缺的工具，通过提升告警处理的效率、增强攻击识别的精确度以及实现响应措施的自动化，显著增强了企业的网络安全防御能力。

4）报告的自动生成

在安全运营工作中需要生成各种各样的安全报告，也是一个比较繁重的工作量，人工智能大模型的文本生成能力在报告的书写方面可以提供大力的帮助，可以高效、优质地生成网络安全需要的各种报告。

3. AI 赋能数据安全

AI 大模型在数据安全方面的应用日益重要，尤其是在数据分类分级和数据脱敏这两个关键环节，它们有助于提高数据管理和保护的效率与精确度：

1）数据分类分级

自动化分类与数据标签生成：AI 大模型可以结合数据字典、建表语句的注释、库表的样例数据学习和理解库表中不同数据的上下文和内容，自动对数据进行分类和分级。这基于大模型对自然语言理解的能力与推理能力，例如，它可以识别出个人信息、财务记录或健康数据等敏感类别，然后按照数据分类分级标准和规则相应地打上不同的安全标签。这样可以确保高敏感度数据得到更严格的安全控制和管理。

2）数据脱敏

虽然已经有各种各样的静态、动态脱敏产品，但数据脱敏一直是一个解决得不够好的问题，AI 大模型在基于对数据内容的理解上，有可能可以做更好的数据脱敏。

- 智能脱敏策略：AI 大模型能够根据数据的分类和分级结果，智能选择最合适的脱敏方法。例如，对于高度敏感数据，模型可能会选择强脱敏策略，如完全替换或随

机化；而对于较低敏感度的数据，则可能采用弱脱敏，如部分遮盖或偏移。这种精细化的策略能够平衡数据保护与业务需求。

- 动态脱敏：AI 技术支持在数据使用过程中实施动态脱敏，即根据用户角色、访问环境和使用场景的不同，自动调整数据的显示方式。例如，对于内部审计人员可能展示更多细节，而对于外部合作伙伴则提供高度脱敏的数据视图，以此来最小化数据泄露的风险。
- 自动化检测与优化：AI 还能持续监控数据处理过程，自动检测数据脱敏的效果和潜在的漏洞，根据反馈优化脱敏算法和策略，确保数据保护措施的有效性和适应性。

通过这些应用，AI 大模型不仅加强了数据安全防护的智能化水平，还提高了数据处理的效率，使得企业在利用大数据的同时，能够更好地保护个人隐私和商业敏感信息。

3）风险评估与策略制定

基于 AI 的分析可以帮助组织动态评估数据的风险等级，依据数据类型、来源、使用频率等因素，自动调整安全策略。例如，AI 模型可以预测特定数据泄露的潜在影响，据此调整访问权限或加密级别，以减轻潜在威胁。

4. 鉴伪与认知安全

AIGC 的迅速崛起，尤其是文生图、文生视频的迅速产业化，使得图片、视频鉴伪、认知安全问题成为各个国家都非常关注的热点问题。

认知安全问题主要指的是通过各种信息传播渠道，尤其是网络和社交媒体，施加的旨在误导、操纵或影响人们认知、决策和行为的安全威胁。这种问题通常与信息作战、虚假信息（假新闻）、心理操纵、欺骗和误导策略等相关，其目的可能包括政治操纵、经济利益、社会混乱或破坏信任等。

认知安全问题的表现形式可能非常多样，包括但不限于：

- 虚假信息和假新闻：发布不真实的信息试图误导公众观点或者掩盖真相。
- 社交媒体操纵：通过伪装成大量正常用户的机器人账号或者雇佣“水军”在社交媒体上散布特定的意见或信息，影响人们的看法。
- 心理操作：利用心理学原理（如恐惧、偏见、归属感等）通过设计的信息策略影响人们的情感和决策。
- 网络钓鱼和诈骗：使用制造的信息或场景误导用户，窃取个人信息或金钱。
- 宣传和虚假旗帜行动：通过各种媒介散布特定政府或组织的政策、观点，有时候也包括伪造事件制造虚假印象的行动。

在数字时代，信息可以迅速传播，假新闻、误导性信息或深度伪造内容（deepfakes）等可以对公众认知造成严重影响，干扰民主过程和公共政策的制定。随着生成式人工智能技术的进步，图像、视频、声音均可伪造并且真假难辨。在俄乌战争中，认知战表现得非常明显，涉及信息操作、心理战术以及社交媒体的广泛使用，目的是影响国内外的公众意见和政策制定。

认知安全领域主要关注的是保护信息系统和网络空间中的知识和信息不受操纵和误导的威胁。这个领域的问题可以从多个角度来考察，包括技术、心理和策略等方面。解决认知安全问题的解决方案复杂多样，包括法律法规的完善、提高公众的信息识别和批判性

思维能力、加强信息源的审核机制、发展和应用技术手段检测及阻止虚假信息的传播等。此外，政府和社会组织的合作，以及国际间的协调和合作也是对抗认知安全威胁的重要方面。技术方面主要依靠人工智能技术，“以魔法打败魔法”。

总的来说，大语言模型的发展为人工智能技术的进步提供了新的动力，同时也对行业参与者提出了新的挑战和责任，推动了技术、伦理和社会问题的深入探讨与解决。

四、 市场分析

（一） 国外安全大模型代表性供应商

1. Anomali

Anomali 是一家 2013 年创立的美国网络安全公司，公司从开发和提供威胁情报产品开始，2013 年推出 ThreatStream 威胁情报平台（TIP）的第一个版本。2016 年，公司更名为 Anomali，并推出了新产品和新的威胁情报方法。包括提供 SaaS 和本地平台，供客户上传日志。它推出了第二个产品 Anomali，后来成为 Anomali Match。2019 年，Anomali 推出了 Anomali Lens，一个网络浏览器扩展，可突出显示并从网页收集相关威胁数据。数据被添加到 ThreatStream 中，并使用 Anomali 的 Match 平台与内部网络事件进行匹配。这是一种企业威胁检测服务，可将数据与现有 IOC 的威胁情报进行匹配。2021 年，Anomali 加入 MITRE Engenuity 的威胁知情防御中心，在攻击流项目上进行合作，以更好地了解对手的行为并提高防御能力。

2022 年 3 月，该公司发布了云原生 XDR（扩展检测和响应）解决方案。它与 Anomali 的威胁情报和 IOC 库合作，帮助公司改进现有的安全基础设施。它可以与 MITRE ATT&CK 框架和其他安全框架集成。2023 年，该公司开始提供由人工智能(AI)驱动的安全分析。

毫无意外地，Anomali 是拿 AI 这个“新锤子”，将自身优势：威胁情报、MITRE ATT@CK 的 TTP、云原生XDR 做成人工智能驱动的网络安全运营平台，通过 AIGC 的分析能力，自动收集威胁数据并驱动检测、优先级排序和分析，从而在几秒钟内实现从检测到修复的安全性。Anomali 安全运营平台由安全分析、精心策划的 GPT 和世界上最大的情

报存储库提供支持，可自动将所有安全遥测数据与主动威胁情报关联起来，以实时阻止攻击和攻击者。

2. Check Point Software Technologies

Checkpoint 是防火墙的发明者，老牌网络安全公司。

Checkpoint 的研究院将 AI 当作顶级优先级在进行研究，在 Quantum Cyber Security Platform Titan Release 中，有三个方面采用 AI 赋能。

- 采用深度学习的高级 DNS 安全；
- 网关式的零钓鱼防护；
- IoT 安全；

研究员现在在研究中的 AI 赋能网络安全的方向：

- 自动化、智能化的访问策略管理；
- 智能化的零信任、基于身份的策略管理；
- 加密流量分析；
- AI 的可解释性；

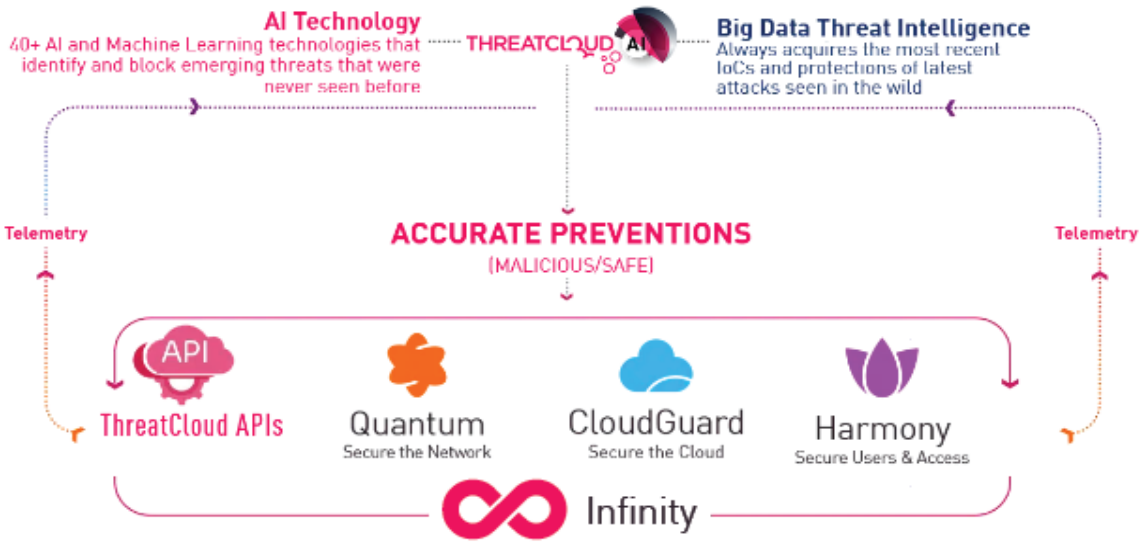


图 3 Infinity 平台中对 AI 的使用

其它产品中，Infinity ThreatCloudAI 被称为 Checkpoint 安全解决方案的中枢神经系统。作为 Infinity 核心服务的一部分，ThreatCloud AI 每天都会汇总和分析大数据遥测数据和数百万个妥协指标(IoC)。Checkpoint 的威胁情报数据库来自 150,000 个连接的网络和数百万个端点设备，以及 Check Point Research (CP)和数十个外部源。超过 50 个引擎配备了基于 AI 的特性和功能。通过 Quantum 增强跨网络、通过 CloudGuard 增强云、通过 Infinity 增强操作以及通过 Harmony 增强用户访问的安全性。Infinity ThreatCloud AI 利用 50 多个 AI 引擎和从数亿个传感器收集的大数据来阻止网络钓鱼、勒索软件、DNS 和恶意软件攻击。阻止网络钓鱼、恶意软件和零日攻击，这些攻击没有可用的 IoC、签名或补丁。IoC 在不到两秒的时间内即可在整个堆栈中共享，包括云、移动、物联网、网络和端点。

2024 年 1 月 30 日，Checkpoint 发布了 Infinity AI Copilot，结合了 AI 和云端技术，致力于提升安全团队的工作效率和效能，以应对全球资安从业人员日益短缺的状况。藉由以 30 年的端到端安全情报进行训练，Infinity AI Copilot 成为安全团队强而有力的后援。透过生成式 AI，Infinity AI Copilot 得以同时作为管理与分析助理，自动执行复杂的安全任务，并主动应对威胁，有效节省处理日常任务的时间，以便安全团队投入策略创新。将 Infinity AI Copilot 无缝整合至 Check Point Infinity 平台，Check Point 在端点、网络、云端等环境提供一致性的安全体验。Infinity AI Copilot 目前已有预览版，并预计于 2024 年第二季度全面推出。

重点特色包括：

- 加速安全管理：Infinity AI Copilot 可以节省多达 90%的安全任务管理工作所需的时间，包括事件分析、实施和故障排除。由于节省了时间，安全专业人员可以将更多时间用于战略创新。
- 管理和部署安全策略：管理、修改和自动部署特定于每个客户策略的访问规则和安全控制。
- 改善事件缓解和响应：利用人工智能进行威胁搜寻、分析和解决。
- 监督所有解决方案和环境：AI Copilot 监督整个 Check Point Infinity 平台上的所有产品（从网络到云再到工作空间），使其成为真正的综合助手。
- 进行简单的自然语言处理：与 Infinity AI Copilot GenAI 交互就像与人对话一样自然。它可以通过任何语言的聊天进行理解和响应，使用户更轻松地沟通和执行任务。这种自然语言能力促进无缝交互和有效的任务执行。在人工智能指南的帮助下进行威胁搜寻、调查和分析事件。编写并运行事件响应手册。

3. Cisco

2023 年 12 月 6 日,思科推出了思科安全人工智能助手 Cisco AI Assistant for Security。这是思科在安全云（思科统一的、人工智能驱动的跨域安全平台）中普及人工智能方面迈出了重要一步。人工智能助手将帮助客户做出明智的决策，增强他们的工具功能并自动执行复杂的任务

防火墙策略人工智能助手：思科安全人工智能助手在思科云交付的防火墙管理中心和思科防御协调器中上线，以解决设置和维护复杂策略和防火墙规则的巨大挑战。管理员现

在可以使用自然语言来发现策略并获取规则建议，消除重复的规则、错误配置的策略和复杂的工作流程，提高可见性并加速故障排除和配置任务。

适用于所有防火墙模型的人工智能加密可见性引擎：当今大多数数据中心流量都是加密的，而无法检查加密流量是一个关键的安全问题。解密检查流量需要大量资源，并且充满操作、隐私和合规性问题。随着 7.4.1 操作系统现已在整个思科安全防火墙系列中可用，客户可以看到 AI 通过加密可见性引擎走得更远。加密可见性引擎利用数十亿个样本（包括沙盒恶意软件样本）来确定加密流量是否正在传输恶意软件。它可以判断流量来自哪个操作系统以及哪个客户端应用程序生成该流量。

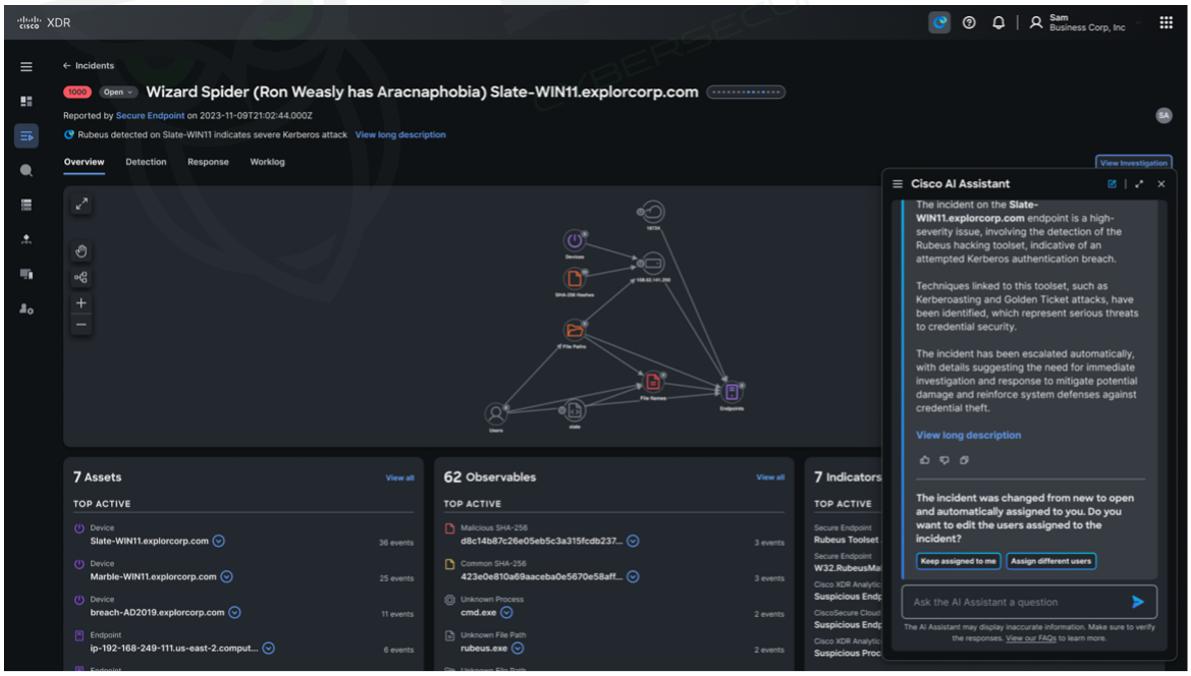


图 4 思科 AI 助手

Cisco Panoptica 专注于成为代码到云的安全解决方案，它提供跨集群和多云环境的无缝可扩展性，并使用户能够获得所需的可见性、上下文和见解，以更好地了解这些环境中最关键的风险。

人工智能助手可以通过日常语言提供即时自定义帮助，帮助用户理解、确定优先级、调查和修复您的特定安全问题。因此，用户可以问“我最重要的弱点是什么？”之类的问题。和“帮助我了解这种攻击路径以及如何修复它。”因此，它对用户的实时环境具有感知和情报，包括 Panoptica 跟踪的有关您的状态、漏洞和攻击路径的所有数据。

Panoptica 的 DSPM 解决方案增强了准确评估所发现的攻击路径的严重性的能力，其细节比以往更加精细，将攻击路径扩展到数据源。这些 DSPM 功能根据内容和启发式提供具有更高程度的优先级和分类的上下文，从而提供对更多服务和数据清单的更深入扫描，帮助团队更好地可视化和评估云中特定资产的敏感性。

4. CrowdStrike

Charlotte AI 是一款面向 CrowdStrike Falcon®平台用户的对话式 AI 助手。它使用行业领先的生成式人工智能技术，使用户能够提出简单语言的问题（“向我展示关键检测结果”、“分散蜘蛛使用什么策略？”等），从而快速呈现 Falcon 平台数据。

目前，Charlotte AI 为用户提供来自 CrowdStrike Falcon® Insight XDR、CrowdStrike Falcon® Spotlight、CrowdStrike Falcon® Discover、CrowdStrike Falcon® Intelligence 模块和 CrowdStrike Detect API 的信息。用户可以使用现有的 Falcon Agent 访问 Charlotte AI,Charlotte AI 目前仅支持英语对话。

Charlotte AI 能够从受支持的 CrowdStrike API 和模块（例如 Falcon Discover，它具有与第三方数据源的连接器）获取和显示数据。Charlotte AI 的底层模型是根据 Falcon 平台文档、Falcon 平台 API 以及 CrowdStrike 的高保真安全遥测进行训练和调整的。

Charlotte AI 未使用有关客户个人身份信息(PII)的信息进行训练，根据 CrowdStrike 的数据处理协议(DPA)，第三方不得在其 LLM 技术开发中使用客户的问题和答案。

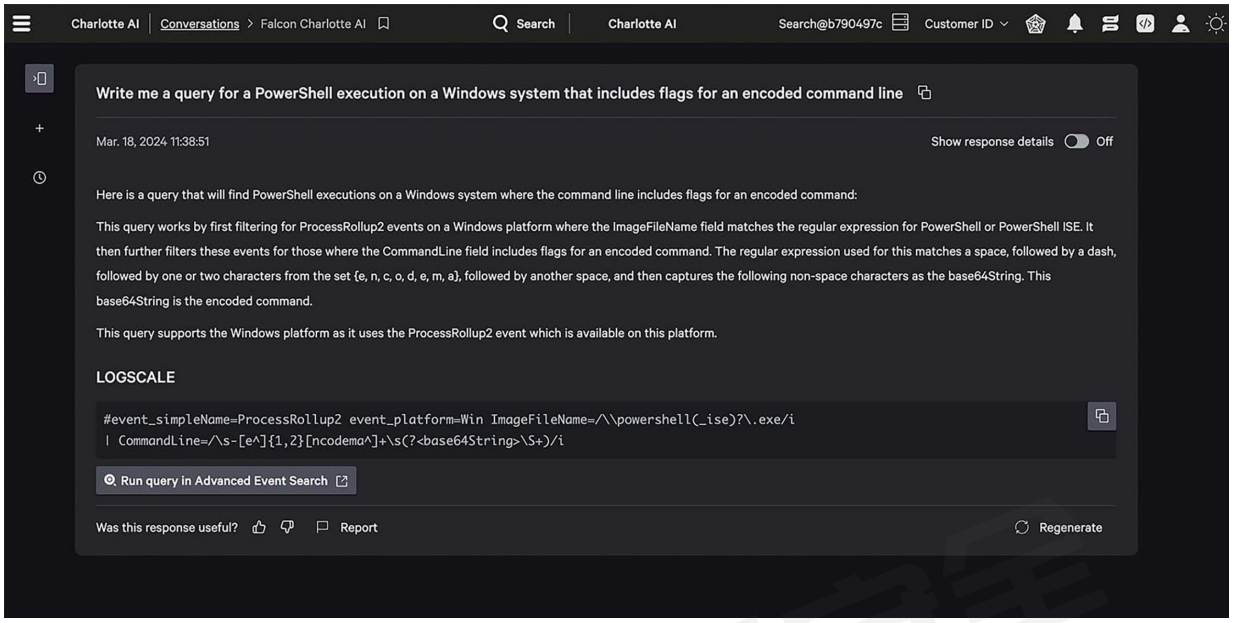


图 5 Charlotte AI 使用示意图

Charlotte AI 的底层架构利用了多种尖端的大语言模型和其他人工智能技术，包括第三方模型和自主研发的人工智能技术。

5. Darktrace

Darktrace 是一家在伦敦股票交易所上市的英国网络安全公司，最新的消息是 Thomas Bravo 将收购 Darktrace 并将其私有化。

Darktrace 认为将人工智能应用到防御技术中预防、检测、响应和恢复的各个环节是有效应对不断变化的威胁形势的必然选择，但并非所有人工智能都是一样的：不同类型人工智能有自己的优点和缺点，适用于不同的网络安全用例。

Darktrace 使用一种称为自学习人工智能的方法，自学习人工智能是一种多层人工智能方法，包括数十种人工智能技术和数百种模型。从本质上讲，Darktrace 的自学习人工

智能以多种无监督机器学习技术为基础，包括神经网络、贝叶斯元分类概率模型、各种聚类技术、大规模计算正则化技术以及许多其他技术以及监督机器学习模型。

Darktrace 有多种网络安全产品：

- Darktrace/Email™，能够立即发现并阻止威胁。
- Darktrace PREVENT™，帮助安全从业者从内到外强化安全。人工智能持续监控攻击面的风险，高影响漏洞和外部威胁。看起来也在环境内部暴露潜在的脆弱性攻击路径和高价值目标。
- Darktrace DETECT™使用 AI 发现威胁通过分析网络中的数千个指标实时并揭示人类不会看到的细微差异。
- Darktrace RESPOND™能够自主地以机器速度采取行动抵御攻击，将响应时间从数小时和数天缩短至仅几秒。
- Darktrace HEAL™可以帮助组织评估他们准备好进行攻击并在现实世界中进行练习场景。在攻击期间，人工智能有助于确定修复的优先顺序增强人类团队的行动。攻击后，人工智能允许企业从网络攻击中恢复并恢复正常状态比人类团队更快、更自信地单独运营。
- Cyber AI Analyst™发现个别异常事件，通过 Darktrace 的核心自学习人工智能，使用监督机器学习将不同的攻击迹象拼凑在一起，然后为人类响应者排好优先级。它产生事件突出显示攻击的每个阶段的即使非技术人员也能理解的自然语言摘要。Cyber AI Analyst 中也有使用大语言模型。

6. Dropzone AI

Dropzone AI 是 2024 年 RSAC 创新沙盒决赛十强之一，口号是：Dropzone AI 复制了精英分析师的技术，并自主调查每个警报，无需剧本，无需代码，无需提示。主打 SOC 的运营自动化。

SOAR 的 Playbook 手册在网络安全运营方面发挥了很大作用，即使有一个很好的拖放界面，复杂性也会很快失去控制。但 DropZone AI 生成式 AI 不使用 if-then 逻辑，而是使用递归推理。生成式 AI 将能够比 Playbook 更进一步推动网络安全警报调查。最后为人类分析师撰写概述上述内容的总结报告，以及作为证据的原始数据的链接彻底的警报调查需要多个步骤，每个步骤都需要推理来评估证据，确定下一步是什么，并创建必要的步骤，在正确的系统（防火墙流量日志、Splunk）中查询以提取所需的数据。

DropZone AI 与其它网络安全大厂不同，没有那么多安全产品，产品以安全智能体形式 SaaS 服务的形式体现，不再是作为安全辅助的聊天形式出现，而是直接对接用户的 SOC 等安全运营平台，与现有产品/服务的整合能力就成了首要考验，以下是 Dropzone 与各类工具的整合能力：

云：已支持 AWS，即将支持 Microsoft Azure、Google Cloud

电子邮件：已支持 Microsoft Exchange、Gmail

终端安全：已支持 CrowdStrike、Microsoft Defender、SentinelOne、Osquery

身份管理：已支持 Okta、Microsoft Active Directory、Microsoft Entra

恶意代码检测：已支持 VirusTotal、CAPA、Hybrid Analysis

网络安全产品：已支持 Palo Alto Networks、Cisco Secure Firewall、Nmap、Tshark、Zeek

生产力工具：已支持、Google Workspace、Slack、Microsoft Office365

SIEM：已支持 Splunk、Gem、Panther，即将支持: Sumo Logic、Chronicle、IBM QRadar

威胁情报：已支持 Blocklist.de IP、AbuseIPDB、Ipinfo.io、GreyNoise、Google Safe Browsing、Host.io、National Vulnerability Database、URLhaus、Shodan、PhishTank、UrlScan.io、Censys

工单系统：已支持 Jira Software、ServiceNow

工装：已支持 Unshorten.Me

漏洞：已支持 Nuclei，即将支持 Tenable

DropZone AI 的创始人 Edward Wu 是华人，但核心团队非常多元化，目前已募得超过 1000 万美金的风险投资。

7. Elastic

Elastic 提供了 Elastic AI Assistant for Security 为安全日志的分析提供辅助，采用聊天机器人形态帮助安全运营人员做告警的研判和辅助分析。

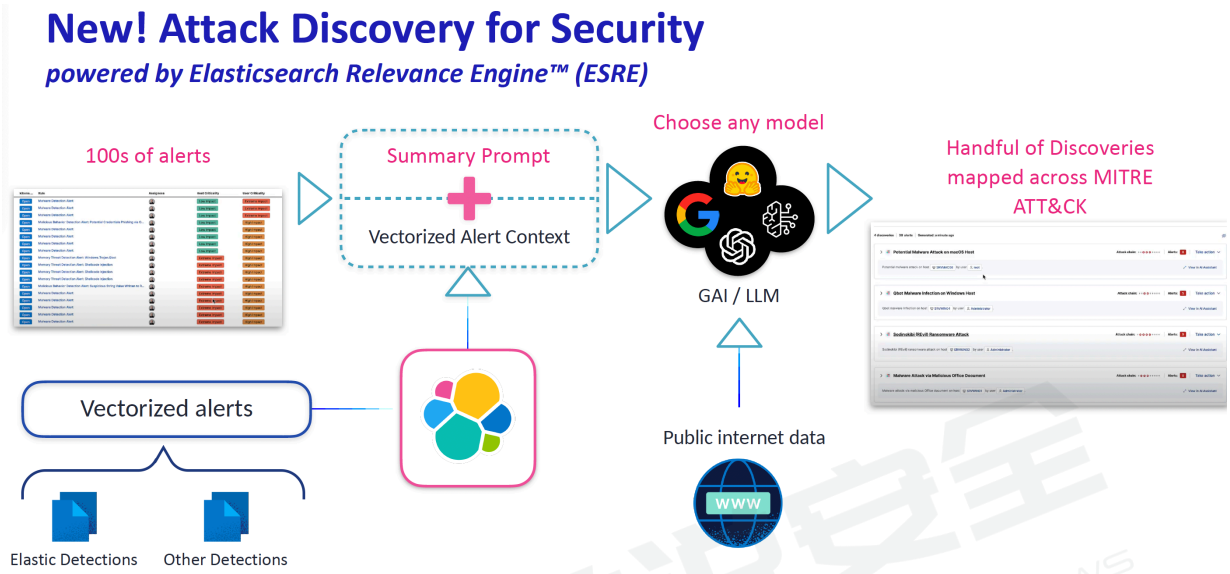


图 6 Elastic ESRE 中对大模型的使用

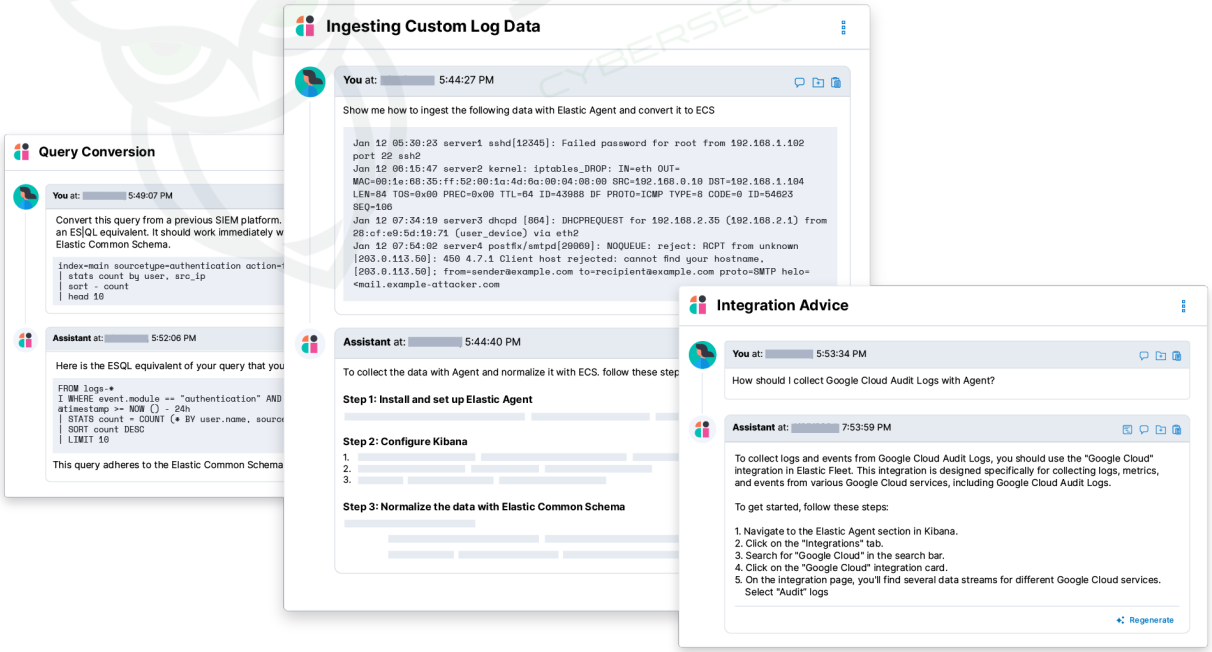


图 7 Elastic 中的对话式 AI 助手

8. Flashpoint

Flashpoint 是一家专业从事网络威胁情报的公司，提供针对网络安全领域的威胁情报、数据分析和研究服务。帮助企业和组织了解、识别和防范网络攻击、欺诈行为以及其

他形式的网络犯罪。Flashpoint 使用数据收集技术和深度分析方法，监控和分析深网（Deep Web）、暗网（Dark Web）以及其他网络空间中的非法活动和威胁信息。

Flashpoint 近期把 AI 大模型技术用于威胁情报分析,包括:用于使用自然语言寻找情报主导洞察的 Ignite AI、用于快速数据分析的 Echosec AI 以及用于跨互联网可扩展识别新数据的自动源发现。通过将人工智能集成到威胁情报运营中，Flashpoint 不仅满足，而且预测客户不断变化的需求，为他们提供抵御网络、物理和地缘政治威胁的强大屏障。

- Ignite AI

Ignite AI 通过从专家制作的广泛情报目录中提取以对话语言查询的情报主导响应，简化了威胁研究的复杂性。它旨在通过易于使用、简洁的、可操作的、准确且值得信赖的情报以及完整的源参考来提高安全从业人员的速度和效率。

- Echosec AI

Echosec AI 在几秒钟内处理数千条社交媒体帖子，增强了及时提供有关重要地点、事件和主题的相关见解的能力。这个工具可以让新手分析师快速获得对社会情绪和趋势的细致了解，提供超越基本数据聚合的深入分析。

- 自动源发现(ASD)

ASD 代表了大规模、快速发现新数据源的创新方法，确保全面的数据覆盖符合客户的情报需求。

9. Fortinet

Fortinet 在网络安全产品中使用各种机器学习和深度学习技术有十多年的历史，如今已经集成了从用于恶意软件检测的十亿多节点人工神经网络到用于警报验证的 Tensor Flow 引擎，为其安全和网络产品组合中的 40 多种解决方案提供支持。

2023 年 Fortinet 推出了一款新的生成式人工智能助手 Fortinet Advisor。在 FortiSIEM 和 FortiSOAR 中首次实施，帮助 SecOps 团队做出更明智的决策，更快、更全面地应对威胁，并简化最复杂的任务。FortiAIOps 简化了 LAN 和 WAN 网络管理，并利用人工智能和机器学习来增强网络操作。FortiAIOps 提供了一种简单易用的方法来管理 Fortinet 网络堆栈（FortiAP、FortiSwitch、FortiGate 和 FortiExtender）。

FortiAIOps 可作为虚拟机使用。在所有部署场景中，它都可以与 Fortinet NOC 和 SOC 工具无缝协作，以统一和简化管理 Fortinet Security Fabric 的方式。

Fortinet 的 AI 产品组合，还包括 FortiGuard AI 安全服务、FortiEDR、FortiNDR 和 FortiAnalyzer。在 Fortinet Security Fabric 中实施的 AI 有助于零日威胁检测，帮助补救复杂的攻击，并使 IT 团队能够在网络和安全问题影响组织之前对其进行改进和解决。

Fortinet Advisor 提供的人工智能大模型为 Fortinet AI 增加了一个新维度，使 SecOps 团队能够直接与 AI 系统交互，以增强威胁检测、分析和响应、生成报告、构建剧本以及修复易受攻击和受感染的系统。

10. Google Cloud

在 2023 年 RSA 大会上，Google 宣布推出 Google Cloud Security AI Workbench，这个平台由专门训练的大语言模型 Sec-PaLM 提供支持。Sec-PaLM 是经过安全用例训练的 PaLM 2 的专门版本，使用 AI 帮助分析和解释潜在恶意脚本的行为，并在极短的时间内检测哪些脚本实际上对个人和组织构成威胁。Sec-PaLM 基于 PaLM 2 模型构造，针对安全用例进行了微调，融合了 Google 强大的安全情报能力，包括 VirusTotal、Mandiant 的对漏洞、恶意软件、威胁指标和行为威胁行为者资料的一线情报。

除了 PaLM 2 大模型之外，Google 还推出了 Gemini 大模型，形成两个大语言模型花开两朵之势。

利用人工智能(AI)和大型语言模型(LLM)的最新进展，Google Cloud Security AI Workbench 着手解决网络安全中的三个最大挑战：威胁过载、繁琐的工具和人才缺口。

- 以对话方式搜索、分析和调查安全数据

人工智能使用户能够更轻松地与其安全事件进行交互并深入研究。只需用自然语言输入问题，AI 就会完成工作。AI 可以生成查询、呈现初始信息，并使得修改和迭代结果成为可能。

此外，人工智能还可以提供安全见解和趋势，汇总并分析来自这些安全事件、实体见解和行为异常的数据，从而为分析师提供加速调查所需的助力。提炼和分析数据成为一种对话式调查体验，可以缩短平均响应时间，但有助于快速确定事件的全部范围。最终，这可以帮助团队更好地利用他们的资源。

- 使用 AI 创建检测

让用户能够使用自然语言提示词让 AI 编写规则。通过迭代用户的提示词，应用风险评分，并进一步完善结果。

- 利用人工智能生成的摘要作出更好、更快的决策

借助 AI，可以帮助用户理解案例和调查中的数据。案例摘要可以自动让用户清楚地了解案例中发生的情况，为您提供有关威胁的指导和解释。迭代结果以达到做出明智决策所需的详细程度

据 Google 的工程师称，AI 提升了其工作效率 5-7 倍。

Google Cloud Security AI Workbench 使用了多个大模型，包括 Sec-PaLM 以及 Gemini 的多个变种。

未来的产品方向：

Gemini in Mandiant 威胁情报：停止威胁；

Gemini in 身份与访问管理

Gemini in 安全控制中心

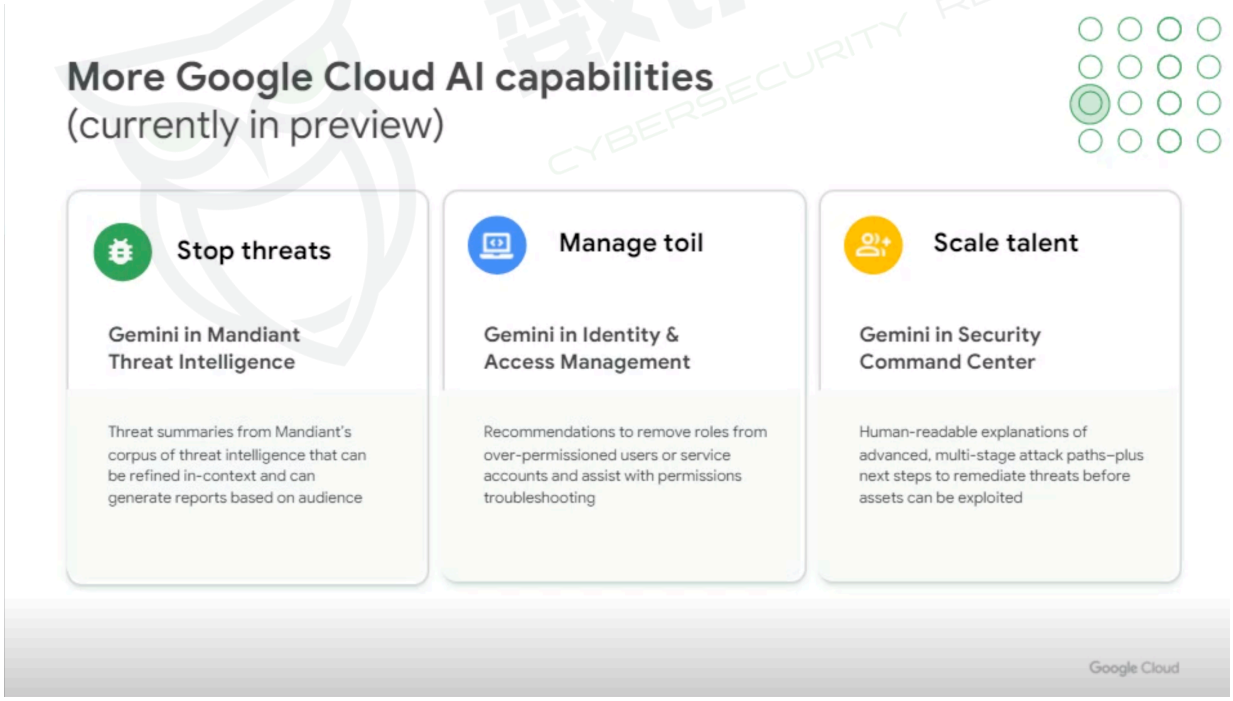


图 8 Google Cloud 中对大模型的应用

11. Microsoft

微软的 Security Copilot 是一个基于 OpenAI 的 LLM 大模型的安全分析和决策支持工具，旨在帮助网络安全专业人员更有效地识别、分析和应对网络威胁。它融合了大量的安全数据和情报，使用机器学习和自然语言理解技术来自动化常规的安全任务，提高响应速度，并帮助安全分析师作出更准确的决策。

Security Copilot 能够处理大量的安全数据，快速对威胁进行分类和优先级排序，帮助团队识别出最紧迫的问题。它还能够提供深度的威胁分析，为安全决策提供数据支持。通过理解和处理自然语言查询，Security Copilot 使得安全专业人士可以通过直接提问的方式获得详细的情报信息和建议，从而大大简化了安全操作过程。

通过整合和自动化安全流程，微软的 Security Copilot 旨在提高安全团队的效率，减少错误和遗漏，同时加强组织的整体安全态势。尽管如此，重要的是要注意任何 AI 驱动的工具都应该在人的监督下使用，以确保决策的正确性和符合组织的安全政策。

在 Microsoft Defender 等产品中，已经添加有“Copilot”按钮，用户可以随时点击“Copilot”按钮，唤出 Copilot 对话框，提供基于当前任务上下文的协助。

目前 Microsoft Security Copilot 已经可以在全球多数国家和地区试用，很遗憾暂未对中国内地与香港开放。

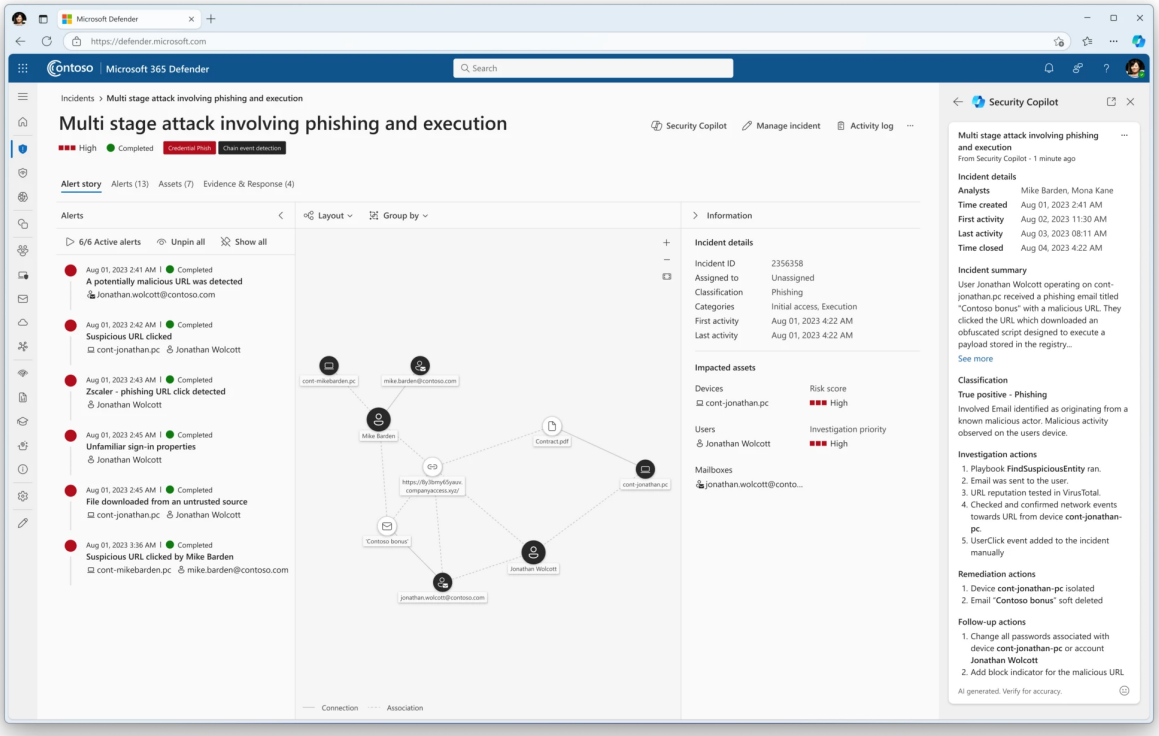


图 9 微软 Security Copilot 集成在 Microsoft 365 Defender 中提供安全报告

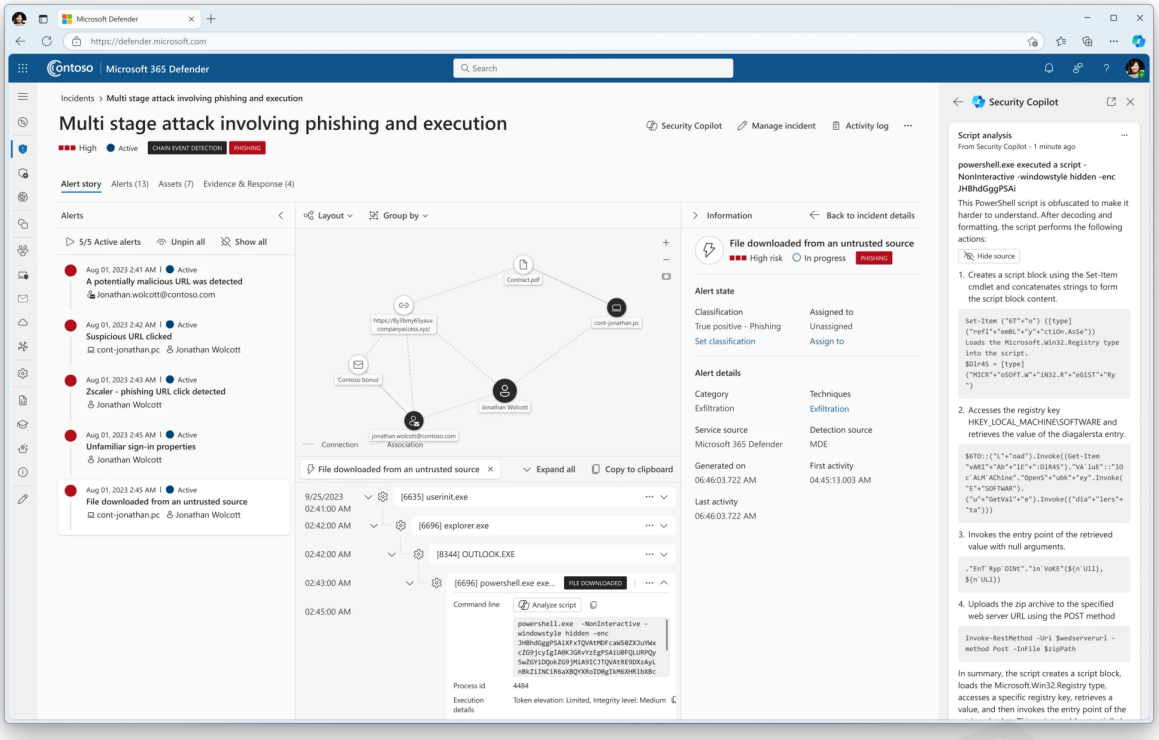


图 10 Microsoft 365 Defender 中嵌入的 Copilot 体验 - 复杂脚本分析和摘要.

12. Palo Alto Networks

Palo Alto Networks 在 2024 年 5 月 7 日一口气发布了三个 Copilot,分别是:Cortex Copilot、PRISMA Copilot、Strata Copilot，背后都是由 Precision AI 加持。

Strata Copilot 可与 Strata 网络安全平台配合使用，包括 SASE 和 NGFW 部署，并通过 Strata Cloud Manager 进行访问。

Prisma Cloud Copilot 支持整个 Prisma Cloud、Code to Cloud™ 平台，协助风险优先级排序、修复、威胁检测和报告。

Cortex Copilot 通过 Cortex XSIAM ®平台提供，可实现更高效、更有效的 SOC。

Copilot 可以:

- 根据用户的环境回答产品知识问题。

- 探索并可视化应用程序、用户和威胁活动。
- 回答有关用户的环境的有针对性的问题。
- 引导配置至最佳状态。
- 搜索有关 IoC 的详细信息（例如，IP 地址、FQDN、URL、域、哈希等）。
- 评估威胁或 CVE 的影响和覆盖范围。
- 通过简单、自然的语言请求来执行复杂的修复操作。
- 协助快速产品导航。
- 打开支持案例，其中会自动收集并主动提交问题详细信息。

Palo Alto Networks 过去的十多年来一直有将机器学习(ML)技术融入到安全技术中的实践，对生成式人工智能AIGC用于网络安全领域是相对比较克制的，首席执行官 Nikesh Arora 在 2023 年 5 月的公司财报电话会议上表示，计划在一年内发布专有的大型安全语言模型，将生成式人工智能嵌入其产品和工作流程中，如今兑现了自己的诺言。

做为三个 Copilot 背后的支撑技术的 Precision AI 融合了传统与最新的 AI 功能：

- 机器学习：十多年来，机器学习技术已内置于 Palo Alto Networks 的许多产品中，通过使用精确的、定义的历史和当前数据作为输入，使安全应用程序能够更加准确地预防、预测和修复安全问题，预测新情况。
- 深度学习：帮助构建预测模型，通过学习大量安全数据来实时预测和检测安全问题。
- 生成式人工智能：使安全工具能够“与人对话”，简化用户体验并总结大量威胁情报。Copilot 建立在 Palo Alto Networks 自己的高度控制的数据集之上，可以缩短解决问题的平均时间（“MTTR”）。

Palo Alto Networks 据称已经有 1,300 多个人工智能模型，用于每天来分析全球数百万个新的遥测对象。每天都会检测到大约 160 万次前一天没有的新的独特攻击，并阻止大约 86 亿次攻击（信息来源：Palo Alto Networks 官网）。

13. Proofprint & Tessian

Proofprint 在 2023 年 12 月份完成对 Tessian 的收购。被收购对象 Tessian 是一家使用人工智能自动检测技术的邮件和 DLP 公司，将 Proofpoint 威胁和 DLP 技术和情报与 Tessian 的 AI 支持的行为检测技术相结合，为组织提供全面防御，以抵御其面对的社会工程学攻击及认为错误，是这次收购的主要诉求。

从勒索软件到 BEC，超过 90%的成功网络攻击以及超过 90%的数据丢失事件（其中 65%的数据丢失事件是由电子邮件发送错误导致的）都是由针对人员的攻击和人为错误造成的。

与其他仅依赖威胁情报或将检测限制于人工智能的解决方案不同，Tessian 的解决方案将能够针对从社会工程到恶意软件再到凭证网络钓鱼的全方位人类威胁发挥很好的有效性，能够利用人工智能对用户活动、行为和数据分类组合进行分析，防止从电子邮件到云再到端点的最关键协作渠道中的数据丢失。

Proofpoint 使用语义分析的投递前大型语言模型(LLM)检测引擎，可以在攻击邮件投递到用户邮箱前即检测出是攻击邮件并进行拦截，在某 500 强客户的对比测试中胜出。

14. SentinelOne

SentinelOne 是一个提供端点安全解决方案的网络安全公司，专注于使用自动化和机器学习技术来提供实时的、全面的防护。它成立于 2013 年，总部位于美国加利福尼亚州的山景城（Mountain View）。SentinelOne 的核心产品是它的端点检测与响应(EDR)平台，旨在自动检测、识别和防止恶意软件和高级持续威胁（APT）。

SentinelOne 的特色在于其人工智能(AI)驱动的行为分析引擎，可以在没有前置文件定义的情况下检测零日漏洞和未知威胁。它不仅能够防止和检测攻击，还能自动响应，包括隔离被感染的系统、终止恶意进程，以及修复由攻击造成的损害。这种自动化和实时反应能力是 SentinelOne 在市场上区别于其他端点安全解决方案的关键因素。

此外，SentinelOne 还提供了丰富的分析和报告功能，为网络安全团队提供了洞察力和可操作的情报，使其能够理解和应对安全威胁。这包括对恶意软件、利用行为、网络通信等方面的深入分析。

随着网络安全威胁的日益增加和变化，SentinelOne 的解决方案适应了迅速发展的网络安全环境，为各种规模的组织提供了一个强大、灵活和高效的安全防护平台。

SentinlOne 在 2023 年推出了 Purple AI（紫色 AI）这款面向现代企业的生成式人工智能驱动的威胁搜寻、分析和响应平台。

其设计初衷是加速安全运营中心(SOC)的进攻策略和响应水平。

在分析客户数据和遥测数据时，SentinelOne 发现其平台中的许多客户查询都非常简单，原因是大多数分析师对他们的角色都是陌生的，并且仍在磨练他们的技能。鉴于这些观察的现实，SentinelOne 开始构建 PurpleAI，目标是把调查过程缩短至 5 到 10 分钟。

利用自然语言总结的威胁结果和人工智能驱动的分析来简化复杂的调查。

通过唯一支持开放网络安全架构框架(OCSF)的 Gen AI 安全分析师获得全面的可视性，以在规范化视图中即时查询本机和合作伙伴数据。

使各个级别的分析师能够使用自然语言查询进行复杂的威胁搜寻。通过建议的上下文后续查询进行更深入的调查，以领先于攻击者。培训分析师根据自然语言提示使用 PowerQuery 翻译进行更复杂的搜索。

通过正在申请专利的搜寻快速入门库进行更快的分析和调查，以减少 MTTD/R 并主动检测风险。通过在已保存和可共享的笔记本中与您的团队合作，促进协作并节省时间。通过人工智能驱动的威胁分析和摘要更快地呈现可操作的见解。

Purple AI 从未接受过客户数据的训练。

客户的流程和见解不会与其他客户共享。

Purple AI 的架构具有最高级别的保障措施。

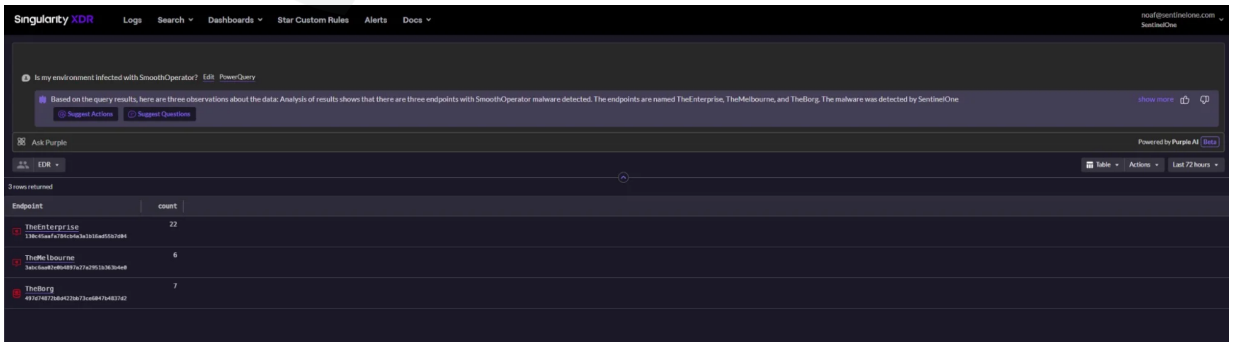


图 11 Sentinel One 的 Purple AI

15. SparkCognition

SparkCognition 是一家人工智能技术初创公司，已经到了 D 轮融资，估值高达 14 亿美金。SparkCognition 致力于开发人工智能驱动的网络物理软件，以确保 IT、OT 和物联

网的安全性和可靠性。该公司的技术能够利用实时传感器数据并不断从中学习，从而制定更准确的风险缓解和预防政策来干预和避免灾难。

SparkCognition 的认知软件 DeepArmor 用于通过先进的机器学习和人工智能来高度准确和高效地预防和检测恶意软件。他们利用屡获殊荣的机器学习技术和专注于国防、工业物联网和金融的专家团队，帮助客户分析复杂数据、增强决策能力并转变人类和工业生产力。

16. Trellix

作为命运多舛的 McAfee 和 FireEye，在被 Thoma Bravo 私有化之后，以 Trellix 的品牌重出江湖，推出 AI 赋能的 Trellix XDR 安全运营平台，采用生成式 AI 技术提高安全运营自动化水平，号称实现 100%告警调查、5 倍提高分析人员工作效率、减少 50%MTTD 和 MTTR 时间。

17. Vectra AI

Vectra AI 提供人工智能驱动的混合攻击检测、调查和响应平台。Vectra AI 平台是为 XDR 提供支持的集成信号，提供跨身份、公共云、SaaS 和数据中心网络的混合攻击面覆盖；AI 驱动的攻击信号情报，实时优先考虑真实攻击；以及集成、自动化和托管响应服务。世界各地的组织依靠 Vectra AI 平台和 MDR 服务以混合攻击者的速度和规模进行行动。

Vectra AI 平台可为您提供发现攻击所需的集成信号，无论您的操作界面如何- Vectra AI、XDR、SIEM、SOAR、EDR。

攻击者没有机会对抗 Attack Signal Intelligence™。Vectra AI 的人工智能像攻击者一样思考，实时自动对真实安全事件进行威胁检测、分类和优先级排序。

18. ZScaler

Zscaler LLM 和人工智能模型直接集成到世界上最大的安全云中，利用每天处理超过 3900 亿笔交易、阻止超过 900 万个威胁和 300 万亿个信号的数据湖。输入大规模、高保真数据和威胁情报，输出经过微调、超感知的人工智能网络安全。

Zscaler 在零信任交换平台和网络产品套件中部署人工智能功能，以识别并阻止攻击链每个阶段的人工智能驱动攻击和传统攻击。

- 攻击面发现：

生成式人工智能使攻击者可以轻松地完成这项一度艰巨的任务，攻击者只需查询与这些资产相关的已知漏洞列表即可。

利用 Zscaler Risk 360 中人工智能驱动的见解，企业可以立即看到这些可发现的（因此有风险的）应用程序和资产（它们的互联网连接攻击面），并将它们隐藏在零信任交换背后的公共互联网之外。

这会立即显着减少企业的攻击面，同时防止攻击者发现薄弱的入口点。

- 攻破

在被攻破阶段，攻击者会利用漏洞来获得对企业系统或应用程序的未经授权的访问。Zscaler AI 创新有助于降低被攻破的风险。

- AI支持的网络钓鱼和 C2 预防

Zscaler AI 模型可检测已知的零患者网络钓鱼站点，以防止凭证盗窃和浏览器利用，并分析流量模式、行为和恶意软件，以实时检测前所未见的命令和控制(C2)基础设施。这些模型结合了威胁情报、ThreatLabz 研究和动态浏览器隔离来检测可疑站点。因此，企业可以更加高效、有效地检测新的网络钓鱼攻击，包括人工智能生成的攻击和 C2。

- 基于文件的人工智能沙箱防御

由 AI 驱动的内联 Zscaler Sandbox 可立即检测恶意文件,同时保持员工的工作效率。传统的沙箱技术让用户在分析文件时等待，或者在第一次允许文件时假设患者零风险。ZScaler 的 AI 即时判决技术可立即识别、隔离和阻止高可信度恶意文件(包括零日威胁)，同时无需等待对这些文件进行分析。这包括通过加密通道（TLS 和 HTTP）和其他文件传输协议传递的威胁。同时，良性文件会被安全、即时地传送。

- 人工智能阻止网络威胁

由 AI 驱动的 Zscaler 浏览器隔离可阻止零日威胁，同时确保员工可以访问正确的站点来完成工作。

- 横向移动

Zscaler AI 功能通过分析用户访问模式并推荐智能应用程序分段策略来限制横向风险，从而减少潜在的攻击爆炸半径。例如，通常可以看到，在 30,000 名有权访问财务应用程序的用户中，只有 200 名实际需要它。Zscaler 可以自动创建一个应用程序段，仅限制这 200 名员工的访问，从而减少威胁行为者的横向移动机会超过 99%。

- 数据泄露

Zscaler AI 开箱即用，可自动发现组织内的所有数据并对其进行分类，使企业能够立即对敏感信息进行分类，同时配置数据丢失防护(DLP)策略，以防止数据因攻击或泄露而离开组织。

（二） 国内安全大模型代表性厂商

1. 360 数字安全集团

360 数字安全集团的安全研究能力是国内的第一阵营，2023 年也积极跟进大模型研发，在 2023 年 ISC 期间发布了 360 安全大模型 1.0 版本，2024 年一季度末发布 360 安全大模型 3.0 版本。

360 的策略是 360 的安全大模型赋能 360 的各种网络安全产品，提高自身各个安全产品的“含 i 量”，暂时没有考虑对外输出 360 安全大模型的能力。360 安全大模型现阶段主要解决网络安全运营与钓鱼邮件防护这两个需求。

360 提出类脑分区协同（CoE）概念，借鉴大脑的功能是分区的，并且是多功能区协同工作的。为此设计了 CoE（Collaboration of Experts）多专家协同安全大模型，与 MoE 类似，CoE 的思想也是基于集成学习（Ensemble Learning）思想，不同之处在于 CoE 中每个专家是专门为一个或多个特定任务训练的，而用户任务是由推理程序与门控模型更为合理地路由一个或多个专家来完成。

360 的安全大模型分为 5 个中枢：

一是语言中枢：实现语言翻译、文本摘要、意图识别能力，可以应用于 Quake 语言、HQL 语言、意图理解、报告生成、指令生成等场景。例如在 Quake 语言中，语言中枢可以帮助安全大模型理解 Quake 查询规则、资产信息、提取信息摘要等。

二是规划中枢：实现程序生成、方案规划、目标拆解能力，可以应用于告警降噪、告警响应、威胁猎杀、攻击溯源。例如，在威胁猎杀场景中，规划中枢规划整个猎杀任务如何进行，拆解成几步等。

三是判别中枢：实现信息抽取、逻辑推理、是非判断、研判检测能力，可以应用于恶意文件判别、恶意邮件判别、EDR 告警判别、攻击流量判别、代码漏洞判别等场景。例如，在恶意邮件判别中，判别中枢通过抽取关联信息、推理多个动作的逻辑，判别是否是恶意邮件。

四是道德中枢：实现情感分析、道德法律能力，对安全大模型从情感、道德、法律维度进行约束，可以应用于内容风控、法律法规、舆情监控场景。例如，在内容风控中，道德中枢对安全大模型的输入输出内容进行检测和过滤，过滤恶意输入，防止生成恶意内容。

五是记忆中枢：实现信息记忆能力，支撑整个安全大模型的信息记忆，可以应用于安全知识问答、私域知识问答、攻防对抗培训等场景。例如，在攻防对抗培训中，提供攻防技战术记忆信息。CoE 的实现机制未知，对现有大模型的体系结构有什么影响也未知。

360 搞安全大模型有三方面的优势：一是拥有大量的网络安全训练数据，包括恶意样本数据、EDR 终端行为数据、网络流量数据、安全运营数据（分别来自 360 信息安全部对 360 数十万台服务、数 T 出口带宽的运营数据）。二是自身人工智能研究院的基础研发能力，360 采购了智谱的 GLM 基座模型源代码并进行了二次开发，对安全大模型的预训练、精调都能提供更好的工具方面的支撑，如 RAG、TAG 工具的支撑。三是拥有数千块 GPU 卡，算力资源虽比不上一线互联网厂商的数万块 GPU 的算力资源，但比传统网络安全厂商的算力还是更强的。

360 信息安全部是 360 安全大模型的第一个用户，“自己的狗食自己吃”是互联网公司的传统，据悉采用 360 安全大模型能实现 MTTR 缩短一半、人均工作效率 30%的提升。迄今为止，已经有至少 5 家客户签约采购了 360 安全的大模型，都是与已有的 360 本地大脑等产品对接做赋能。

图 12 360 安全智能体框架

360 安全大模型的专职研发团队有数十人，数据工程、大模型基础能力等由其他团队提供支撑。

2. 安恒信息

安恒在基于深度学习技术做安全检测方面有几年时间的积累，也是从 2023 年之后开始跟进大模型在安全检测、安全运营及数据安全领域的应用，现在 AI 战略是安恒 2.0 战略的核心，让安恒现在的安全产品 AI 化，是安恒试图寻找的第二增长曲线。

首先，安恒开发了“小恒智聊”插件，提供助手级别的安全辅助工具。把各种基本的安全能力以及好集成的能力都封装到了插件里面，成为一个类 SDK 一样的东西，可以很容易的插到各个产品里面，只要把几个接口一对接就可以用了。甚至发布了小程序的一个版本。小恒智聊插件包括客户与公司一线营销、行销、售前的工作人员也都在使用。

对安全态势感知与安全托管服务，AI 的集成就要深很多，在 AiLPHA 平台上，点击 AI 图标，可以看到高危的告警已经被检索出来。点击一条告警信息，可以看到，经过编码的告警报文。过去，如果要分析这些报文需要攻防专家有非常深厚的攻防基础知识，才能胜任这项工作。现在 AI 引擎通过学习历史的攻防数据安全知识库就能实现 AI 解读报文的功能，该功能可以辅助我们现场的工程师完成告警分析的工作。

数据安全领域是 AI 大模型应用的另一个领域，在数据分类分级工作上，AI 大模型可以提升工作效率达 10 倍以上。通过 AI 大模型进行 API 风险监测是数据安全领域中的另一应用。

3. 金睛云华

金睛云华公司是一家以 AI 技术为核心,致力于将人工智能技术落地到网络安全领域，发现传统手段无法发现的未知高级威胁的网络安全创业公司。公司成立于 2016 年，创始人曲武是人工智能方向的博士，曾在清华大学 KEG 实验室从事博士后研究。先后在启明星辰核心技术研究院和华为安全产品线工作，安全功底深厚。

该公司的业务包括但不限于云鉴高级威胁检测系统(ATD)、云图大数据安全分析系统(CIC)、大数据威胁情报系统(CTI)、云踪全流量威胁取证系统(TFS)、云晰加密流量分析系统（ETD）等。从公司成立之初，就确定 AI 驱动安全为主要技术路线，以深度学习技术为主要技术打造一系列安全产品。包括把程序样本转换成灰度图像，利用深度学习的图像识别的能力来判别程序是否是恶意文件等方法。以及采用 LSTM 技术通过加密通信流量的特征：上下行流量的分布、中间间隔时长等行为特征来判别通信的主体等。建立了 20-30 个“小模型”来解决威胁检测问题，取得了不错的效果。

2023 年初，随着 ChatGPT 为代表的生成式 AI 所展现出的令人惊艳的进步，金睛云华也开始尝试使用生成式 AI 来解决网络安全问题。因此首先尝试的是用生成式 AI 这把“新锤子”，把恶意代码检测等“老钉子”再砸一遍，BERT 是他们用的这把“新锤子”，引入“程序语言”概念，进行预训练和精调之后，取得了不错的效果，形成了金睛云华的检测大模型。

稍后，公司也开始了“助手”类安全运营产品的研发，采用的则是 GPT 的路线，尝试过多个基座大模型，包括 GLM、LlaMa-2、MISTRAL、通义千问等，发现各个模型都有自己擅长的和不擅长的方向，决定采用 MoE 混合专家模型，根据用户的不同任务动态调度多个模型中的一个或多个来完成任务，并形成了金睛云华的运营大模型。

云智（BOC）集成了金睛云华的检测大模型和运营大模型，并结合外围的链，agent 等相关技术，具备安全产品，三方数据，情报，安全工具等的对接和调度能力，具备安全智能体的雏形。BOC 是个相对比较独立的产品模块，可以赋能金睛云华的产品，也可以方便被其它安全公司的产品整合，为其它公司的安全产品提供智能化支撑能力。

在 BOC 赋能的基础上，云鉴 Pro、云踪 Pro、云晰 Pro 就是升级之后的威胁检测、全流量威胁取证、加密流量分析系统。云图 Pro 是采用大模型技术加持的安全态势感知系统。这几个产品的特点是用 BERT 检测大模型增强检测能力，并用 GPT 运营大模型增强智能处理能力，大幅降低了人的工作量，取得很好的效果。

某银行在进行产品测试的时候，拿了 10,356 条 WAF 告警日志来做测试，用 8 张英伟达 4090GPU 的算力，CIC Pro 耗时 12 分 16 秒产生了 248 条告警，自动化运营分析 5 分 21 秒分析后认为有效告警 41 条，对比的人工分析 1 人周，发现 32 条有效告警，用大模型直接分析，耗时 3 个多小时，发现 47 条告警，人工检查，发现有 2 条误报。

金睛云华的发展方向，是采用大模型，聚焦于提高威胁检测能力、提高网络安全运营效率。

金睛云华的产品已经在 2024 年 4 月中旬开始在用户环境进行 POC，首批 POC 客户包括大型金融机构及政府大数据局。

在算力资源上，金睛云华在大庆有自己的数据中心，80 个机柜，400 余张 GPU 卡进行训练和推理。

4. 海云安

海云安是开发安全厂商，开发者安全智能助手 D10 是基于海云安公司的开发安全实践经验打造出的一款开发者安全助手产品，D10 将开发安全实践经验、SAST（源代码检测分析）和 SCA（软件成分分析）技术与当前热门的人工智能 AI 大语言模型技术进行深度融合，将代码安全、合规、质量的检测，自动补全代码，AI 降低检测结果误报率，AI 一键生成修复代码，AI 智能交互问答对话等场景融于一体，帮助开发者在开发编码阶段，从安全、合规、质量、效能四个方面进行全面赋能，助力企业整体研发效能提升。

5. 华清未央

华清未央是一家学术背景很强的安全公司，创始人曾经带队参加 2014-2016 年 DARPA Cyber Grand Challenge 并且取得过不错的成绩，其提出的 Machine Language Model 是全球首个机器语言大模型。

MLM 全面支持对多指令集、多操作系统、多文件格式的可执行程序/二进制程序/闭源软件进行结构分析、语义理解、安全分析。MLM 不依赖软件源代码或调试信息，解决二

进制程序信息缺失和语义理解困难等难题，突破软件分析技术多年来面临的瓶颈。MLM 为逆向工程、漏洞挖掘、恶意代码分析、供应链分析、版权保护、性能优化、生态迁移等软件问题提供高效智能化解决方案。

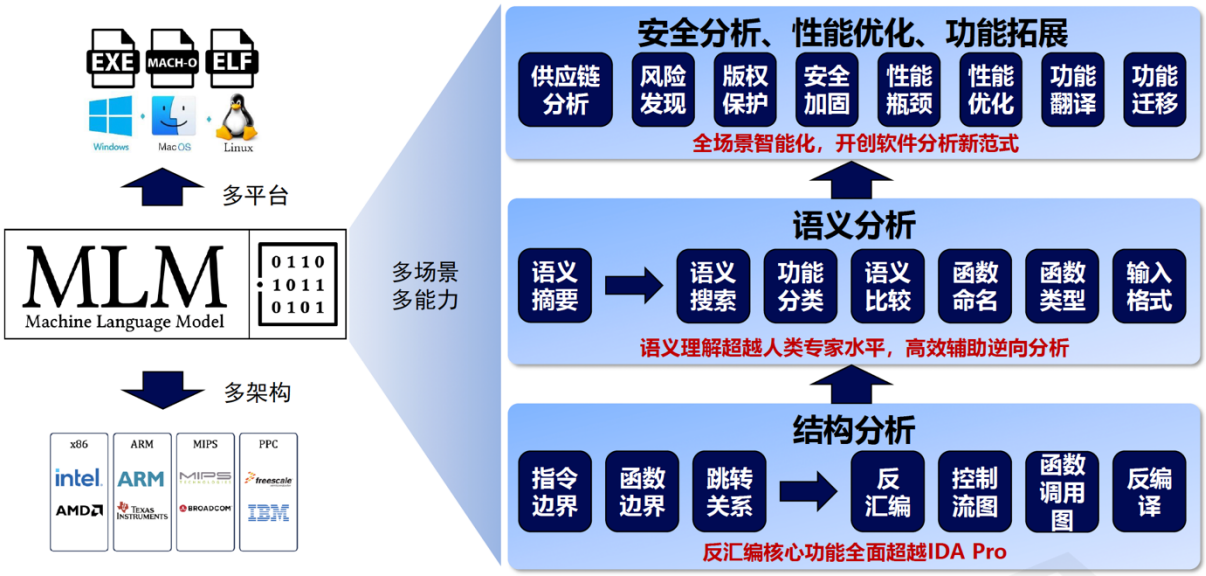


图 13 MLM 示意图

MLM 在软件开发和软件分析领域可以有极其广泛的应用，有可能改变软件世界的生态和形态。目前，华清未央 SaaS 平台已向近万名用户提供服务，并为数十家机构客户完成本地化部署。

华清未央共有六大产线，实现软件开发及安全分析全覆盖：

- 软件逆向分析平台

软件逆向分析平台旨在深度分析和理解软件的结构和语义，形成了以机器语言大模型为基础的软件逆向分析系列工具。

能够无需源码理解各类软件语义，输出深层次的软件分析结果。帮助软件开发人员、安全分析人员深度分析各类形态的软件，理解代码语义，发现安全风险等。软件逆向分析平台提供了一系列功能，涵盖二进制代码结构分析、语义分析及安全分析等，可以满足用

户对二进制代码分析的不同需求。相比于国外同类软件，该平台分析结果更准确，功能更全面，填补国际空白。

- 软件供应链分析平台

软件供应链平台旨在深入溯源软件组件成分并识别风险。平台基于机器语言大模型和海量开源及闭源软件数据，能够有效溯源（无源码）软件的开源组件和闭源组件，识别软件供应链中的潜在供应风险和后门、漏洞等安全威胁，输出全面的供应链安全报告。在供应链依赖关系、组件安全性、许可证合规性等方面表现优异，为关键基础设施行业和各类企业用户提供全面的供应链安全保障。相比于同类分析工具，该平台可以分析二进制软件供应链，可以识别闭源组件，具有更高的识别准确率。

- 软件版权保护分析平台

软件版权保护分析平台旨在保护软件知识产权及取证侵权行为，形成软件版权保护分析系列工具。该平台可以为软件提供防破解保护，同时依托机器语言大模型优异的表征学习能力，高效地提取软件代码语义并从语义层面检测软件代码间的相似性，解决了软件侵权取证场景下无法获取源代码的难题。在代码克隆检测、软件侵权检测等方面具有显著效果，为软件知识产权提供保护。

- 软件生态迁移系统

软件生态迁移系统旨在为用户提供高效、可靠的软件迁移解决方案。该系统基于机器语言大模型的代码理解及翻译能力，能够在不同架构平台、系统环境、编程语言之间实现软件无缝跨生态迁移，确保迁移后的软件具备稳定运行的特性。系统在代码转换、代码恢复、代码补全等方面具有卓越表现，为企业提供全面的软件生态迁移支持。相比于传统的

软件生态迁移服务，该系统极大降低了对于开发人员的依赖，支持无人维护代码的迁移，极大降低了迁移成本。

- 软件漏洞挖掘平台

软件漏洞挖掘平台旨在高效发现和消除软件潜在安全风险，形成了面向各类软件和代码的漏洞挖掘系列工具。该平台支持多节点、多框架的异构漏洞挖掘，能够显著减少人工挖掘的时间和成本，提高漏洞挖掘的覆盖率和准确性。平台提供了全面的漏洞挖掘功能，包括模糊测试、漏洞复现、漏洞去重、POC 生成、漏洞能力分析、漏洞扫描、漏洞修复建议生成等，可以满足用户对 DevSecOps、漏洞检测和管理的多样化需求。相比于传统的模糊测试等漏洞挖掘工具，该平台漏洞挖掘效率更高，而且支持二进制代码静态检测。

- 恶意代码分析平台

恶意代码分析平台旨在为用户提供高效、精准的恶意代码分析与检测解决方案。该平台依托于机器语言大模型，能够自动扫描和分析软件代码，理解代码语义和意图，识别潜在的恶意行为及其代码片段，并输出详尽的检测报告。在后门、病毒、僵尸网络、木马、蠕虫、挖矿软件等恶意代码分析与检测方面表现出色，通过智能化的恶意软件自动分析,该平台能够显著减少人工分析的时间和成本,提高检测的覆盖率和准确性。

6. 华为

华为在 AI for Security 领域进行了长期且持续的投资，从 2013 年开始在德国慕尼黑建立 AI for Security 的研究团队，利用机器学习和深度学习来进行安全流量和文件检测的研究，并转换研究成果到 EDR、Firewall、DDoS、IPS、HiSec Insight、SecoManger、云服务等安全产品中，提升检测和运营的效率，有效解决用户面临的实际问题。

针对人工智能时代网络安全技术所面临的挑战，华为推出了星河 AI 网络安全解决方案，该方案采用云端、网络、边缘和终端四层架构，具备“云端威胁秒级处置、网侧超大规模组网、边界领先检测性能、终端精准勒索防护”安全能力。在 2024HAS 期间，华为发布 L4 级 AI 安全智能体（网络安全高度自主防御），全面进入智能防御时代，实现自主防御新型针对性攻击。

华为具备全栈大模型能力，包括从 AI 芯片（昇腾系列），CANN 计算架构，MindSpore 深度学习框架，MindStudio 开发工具到智算训练网络以及盘古通用大模型。华为安全团队结合在网络安全领域的多年积累，在传统机器学习和深度学习的基础上，进一步增强了生成式 AI 的能力，构建了一个大小模型融合的统一安全大模型。这一模型旨在解决安全检测和运营中的挑战，并已集成到华为的“乾坤云”安全托管服务中，与 EDR 产品、乾坤云安全网关共同为用户提供服务。通过与合作伙伴和客户的共同努力，华为不断推动大模型在安全运营领域的应用和成熟。并逐步延伸到线下，给大客户提供 On-Premise 版本的安全运营大模型能力。

此外，华为还积极探索 AI 在安全领域的多种应用，包括端侧 AI 能力。面对日益猖獗的勒索软件和数据泄露攻击，华为构建了基于行为的 AI 检测能力，以防御未知恶意软件的攻击。华为利用自身的嵌入式 AI 平台 SiteAI 和硬件加速能力，打造了高性能的边缘 AI 检测能力，将 AI 的实时检测能力扩展到安全边缘，为客户提供实时的威胁检测和阻断能力。在统一运维方面，华为利用自身在安全运营方面的丰富经验，将经验和规则转化为 AI 知识，进一步赋能安全运营大模型。华为基于自研多流 ECA 专利算法+ AI 大模型自学习引擎，实现加密流量不解密检测，准确发现藏匿在加密流量中的威胁和攻击流量在嵌入式设备方面。

华为目前聚焦安全检测和安全运营两大主流场景，服务多家大客户，并获得客户好评，以某头部公司为例，每天产生约 100 亿+条日志，10 万+告警，1000+安全事件，按每个安全分析专家每天最多能处理 10 条安全事件评估，需求投入大量的人力资源。借助华为安全运营大模型的能力，目前实现自动处置 90%以上的安全事件,极大降低人工工作量，提升网络安全运维效率。

7. 火山引擎

火山引擎是字节跳动旗下云服务公司，安全业务的商业化也在火山引擎孵化。

火山引擎在大模型赋能网络安全、大模型赋能数据安全方面都有尝试，都是首先从解决自身业务运营中的问题作为出发点。在大模型赋能数据安全场景中，基于大模型的数据分类分级与数据脱敏，是火山引擎认为成熟度比较高，可以商业化的产品。在大模型赋能网络安全场景中，火山引擎认为其能力与火山引擎自身雇佣的水平相对较高的网络安全团队相比还有差距，暂时不具备商业化条件，

火山引擎所采用的大模型底座是字节跳动的“云雀大模型”。

8. 酷德啄木鸟

酷德啄木鸟开发了 Codepecker 专属的代码审计助手“Codepecker Audit Assistant”，简称代码审计助手。审计助手可以理解代码的上下文和逻辑，协助发现和审计代码中潜在的缺陷，如质量错误、安全漏洞、违禁规则等：

代码审计助手基于对历史审计数据的分析，并结合大模型（LLM）进行微调。具体步骤包括：

数据收集：收集并整理大量历史代码审计数据，包括各种编程语言和框架的代码样本及其对应的审计报告。

模型训练：使用大语言模型（LLM）进行初步训练，使模型具备基础的代码理解和审计能力。

微调优化：基于收集到的历史审计数据，对模型进行微调，提升其在不同编程语言和框架下的审计准确性和效率。

测试验证：在实际项目中进行测试，验证模型的审计效果和修复建议的准确性，并根据反馈不断优化模型。

这种基于历史数据和大模型的微调过程，确保了代码审计助手具备高效、准确的审计能力，并能为开发人员提供切实可行的修复建议。

代码审计助手可以显著提高了代码审计的效率和效果。其主要优势包括：

减少人工干预：通过智能分析和修复建议，减少了人工参与和干预，提升审计效率。

提升审计质量：利用大语言模型（LLM）处理能力，可以快速地对大量代码审计结果进行分析和评估，确保高质量的审计结果。

节省时间和精力：自动化审计流程节省了审计人员的时间和精力，使他们能够专注于更高层次的工作，如复杂漏洞的深度分析和修复策略的制定。

通过这些功能，代码审计助手帮助开发团队更高效地发现和修复代码中的问题，提升软件的安全性和可靠性，为企业和用户提供更有保障的软件产品。

9. 灵云数科

灵云数科是一家极具创新活力的网络安全初创企业，创始人白应东自 2022 年初带领团队从邮件安全这个长久以来一直没有得到很好解决的痛点切入安全市场。

中国的邮件系统生态与西方国家有很大差别,在政府和国企中 SaaS 方式采用的较少，本地化部署较多，邮件安全的问题直接采用西式 SaaS 服务来解决的可能性也小很多。

灵云数科敏锐的察觉到人工智能大模型所带来的机会，同时采用 Bert 与 GPT 两种技术路线在邮件安全上，取得了不错的成绩。并且在模型共享方面做了一些创新，使得不同的组织能够共享邮件安全的一些检测模型，但又不会泄露企业的机密信息。目前灵云数科的邮件安全产品已经有数百家机构采用，多数是关键信息基础设施机构。

灵云数科不甘心只做邮件安全，在尝试将人工智能大模型技术应用在更广泛的领域，比如社会工程学攻击的防范等。

10. 绿盟科技

绿盟是中国最早成立的网络安全公司之一，以攻防技术见长。抗 D、扫描器、WAF 曾经是其王牌的三个产品。绿盟从 2009 年开始尝试采用机器学习等技术用于防火墙、WAF、IPS/IDS 等“墙”类产品的赋能，2015 年成立天枢实验室，聚焦在 AI 方向的研究，最大的成就就是形成了一个相当大规模的网络攻防知识图谱，包括漏洞、攻击工具、组织、技战法等，这个庞大的知识图谱在绿盟安全大模型的训练中起到了关键作用。大模型赋能的渗透测试工具是绿盟的一大特色，与其它友商形成了明显的差异化。

2019 年绿盟推出第一款 AI 赋能网络安全运营的产品，2023 年开始大模型赋能网络安全的研究，历经 5 个月研发，2023 年 9 月推出第一版的大模型安全运营工具。

在进行产品 AI 赋能方面，绿盟首先赋能的是 SOC 类产品，包括本地 SOC 和云端 SOC 平台，计划中的是做攻防类产品的 AI 赋能以及数据安全产品的 AI 赋能。

11. 奇安信

奇安信是国内头部的网络安全厂商，产品线非常长，历史上也有冬奥会网络安全保障零事故等辉煌的战绩。

奇安信 2023 年下半年推出了 QAX-GPT，定位在智能研判、智能调查和智能任务，也就是智能处置，达到促进网络安全运营升级、提高告警研判效率、增加产能的目的。

奇安信在大模型上的技术积累说起来是个“无心栽柳”的故事：2019 年开始想用人工智能解决天眼的客户服务问题，做了一个“天眼小助手”，把天眼产品的一些文档，日常一线二线服务工程师问的一些问题都给整合起来，然后通过蓝信的入口或者微信的入口，让用户提问，然后由 AI 引擎直接给回答。这件工作投入了较大的力量，包括数据处理与研发精力，但积累下来的经验在大模型兴起的时候就用上了。捎带着还在国际比赛和国内的比赛中获得了不错的成绩。

QAX-GPT 的策略，是先与奇安信的优势产品“天眼”做对接，做流量侧告警研判，并且有能力对接其它厂商，如中睿天下、绿盟、安恒的流量分析设备。然后再对接 SOC 类产品，最后再对接天擎、BAS、椒图的主机安全类产品等。

智能研判：原来大部分网络安全运营人员的精力都聚焦在做告警的监测研判这些事情上。经过机器人的整体的一个混编，包括升级之后。那大量的安全人员的精力就可以从监测和研判上抽出来，大部分精力就可以聚焦在高难度的问题解决上。整体的产出会变多。机器处理速度快，又是不眠不休，“一台机器人等于 60 个安全专家”就是这么来的。

智能调查：调查过程需要与其它部门配合，会遇到对方配合度不高、配合不够及时等问题，不仅工作进度受影响，还可能会引入一些负面情绪。安全机器人通过蓝信自动与相关人员确认威胁，可以减少沟通成本。与天眼、天擎、椒图、SOC 等多源数据的关联溯源分析也可自动完成，不需要再登录各个系统来完成工作。当然，最大的好处，还是可以与安全运营人员进行自然语言的交流。

智能任务：核心点是 QAX-GPT 可以直接生成一个可执行任务，也可以再做一些分配，然后直接同步。任务类型，现在是梳理了 11 个大的场景。就是相当于说先有一个任务的一个大概的描述，这就是通过机器人，然后通过模型直接根据下面的一些待处置的任务生成的一个综述。

奇安信的 QAX-GPT 以“安全机器人”形式出现，可以是独立部署的软件，或者以软硬件一体的形式提供给用户。三种交付形态，第一种是纯粹的裸模型提供给客户。他想怎么用就怎么用？第二种就是把模型装在安全机器人里面，直接提供出去。最后一种，就是直接把整个模型生产的流程、整个环境交付给客户，包括数据标注的一些平台，包括算力平台，它可以直接在上面去做那个模型的一些训练。目前客户大多选第二种交付模型。

奇安信用于安全大模型训练的有 1,000 多张显卡，近百人的技术团队持续研发，在整理训练数据阶段动用了公司包括安全服务人员在内的更多的资源。

作为副产品，Q-GPT 在奇安信 HR 的简历筛选、工程师的编程辅助（代码完成、代码缺陷发现）方面也已经可以发挥作用。

12. 深信服

深信服是国内网安巨头中比较另类的一家公司，以标准化产品、渠道销售为主，并且网络安全产品与云计算产品、数通类产品齐头并进。

其在安全大模型方面的投入也是最坚决的，据说为安全大模型的研发停掉了其它产品的研发，组织了 400 人团队，500 张 A100/A800 显卡集群。2022 年 12 月底开始训练、2023 年 5 月 18 发布第一个版本，2023 年 9 月迭代了 2.0 版本，在 2024 年 1 月迭代 3.0 版本。目前，安全 GPT 已发布三个大革新，分别提供基于自然语言的安全运营智能助手，模型自主值守研判，针对高对抗流量攻击的检测解读，以及针对钓鱼攻击的检测研判与处置能力。

深信服自身有 MSSP/MDR 业务，其安全运营中心自然就成了第一个实验场，这对大模型来说是非常重要的事情，可以及时得到反馈数据，实现所谓的“数据飞轮”。

据称，在多次实战攻防演练以及头部客户测试中，深信服安全 GPT 针对高对抗、高绕过攻击，比如 web 0day 攻击，均表现了远超传统安全设备的检测能力。

深信服的大模型也有两个应用场景：安全运营与安全检测。

安全检测：又细分为网络流量攻击检测与钓鱼邮件检测。其中网络流量攻击检测是通过海量的 HTTP 流量、日志、代码等数据的预训练而成的深信服大模型具备了 HTTP 流量理解能力、代码理解能力、攻防对抗理解能力和安全常识理解能力,类似一个攻防专家。

安全运营：提供对话式的“辅助驾驶”模式，为安全运营人员提供图文并茂的安全态势解读，为安全运营人员提供处置建议。同时，也应一些用户的要求，提供了更为激进的“自动驾驶”模式，安全 GPT 分析完成后，即可对划分为自动化处置的问题自动产生动作，待人员决策事件留给运营人员处理。

数据安全方向，大模型的应用已经预研了大半年的时间，可能 2024 年会推出数据分级分类的大模型，然后是访问控制的大模型。

据深信服自己声称，目前已经有超过 150 多家客户部署了深信服安全 GPT，在走访深信服安全 GPT 客户的过程中了解到，2023 年首批采购深信服安全 GPT 的客户，深信服派出了几十人的团队帮助客户做安全 GPT 与其它安全系统乃至业务系统的对接，以实现安全处置的闭环，客户对深信服的响应速度非常满意。

13. 腾讯

腾讯的安全大模型由 CSIG 下的多个团队分别开发，其中最为成熟的是腾讯数据安全大模型。

腾讯用混元大模型的底座，以及积累的丰富的数据安全威胁情报、数据治理经验，为大模型的训练提供支持。有专门团队长期研究 AI 在数据安全领域的应用，积累算法和落地经验。通过大模型在数据安全四大领域的应用，包括：敏感数据发现、模型训练数据标注、数据防泄露和溯源等，提高数据保护效率。

首先，在资产清查阶段，利用 AI 帮助发现 API 资产，提高识别准确率。

第二，在敏感数据识别方面，通过学习提高识别准确率。

第三，在风险监控阶段，利用 AI 模型进行持续风险检测，提高输出准确率。

最后，在数据安全分析方面，基于行业大量样本数据训练行业数据模型，帮助识别敏感信息并制定保护策略。

- 数据分类分级的 AI 应用

一种基于 AI 的数据分类分级方法。该方法将数据分为格式属性、集合内不同字段组合、语义内容三个维度。在处理过程中，AI 会根据语义属性进行分类，结合上下文环境，形成更精准的数据属性集合。在规则引擎方面，AI 提供了大量帮助，辅助识别规则格式，以提高数据分类分级的准确性。整个方法与传统规则引擎相互衔接和提升。

- 基于 AI 的敏感数据识别

基于 AI 的敏感数据识别产品及其应用场景。通过微调模型将文本语义信息化成向量化表示，并通过聚类算法自动发现文本之间的语义相似度。在 2023 年 10 月份发布的基于 AI 的敏感数据识别产品,除了传统的基于数据的建模和规则引擎外,还叠加了 AI 能力。该产品主要应用于数据分类分级、数据防泄露和敏感数据追溯等场景，如互联网、金融、汽车、政务等行业。目前，该产品的准确率达到了 99%以上。

- 数据流转风险识别与 AI 应用

针对数据流转过程中的风险行为，通过 AI 模型来识别数据访问行为，梳理数据访问路径。AI 模型在资产相似度、行为相似度和前后上下文关联分析方面的应用，可以帮助用户快速发现资产并扩大范围，优化敏感识别的准确度，精准匹配行业，提高溯源效率。此外，AI 模型还可以在企业内部进行私有化部署，通过调优补充数据模型。

- AI 在数据标注与识别中的应用

首先，通过训练 AI 命名实体识别模型，提高了准确率，误报率有所降低。其次，利用大模型辅助数据标注，大大提高了标注效率，降低了成本。在训练过程中，发现需要大量有标注的数据，而标注人员往往缺乏背景知识。为了提高准确率和节省人力，采用了大模型进行训练。此外，通过大模型辅助数据标注，可以解决简单内容键值标注的问题，从不同维度对数据进行多轮标注，最终产生一段话式的标签。

- 数据防泄露产品的应用与优化

传统数据防泄露产品的真实场景应用与实验室场景存在差距，导致数据防泄露产品使用效果不佳，而 AI 技术可以提高效率、准确率和运营效率。在数据安全场景下，事件特征异常较小，因此明确处理是效率最高的。

14. 天融信

天融信是国内首家网络安全企业，成立于 1995 年。

天融信 2014 年成立 AI 研究团队，开启网络安全领域的机器学习、深度学习、自然语言处理等人工智能技术研究和应用。2016 年，天融信推出基于机器学习技术的恶意样本检测、隐秘隧道检测，2018 年将自然语言理解、NER 技术应用于威胁情报及知识库，2020 年开始研究类大模型，应用类大模型技术的威胁情报与知识库，在精准度、产出效率方面均大幅提升。2023 年，天融信发布天问大模型，2024 年，在下一代可信网络安全架构（NGTNA2.0）下，将 AI 技术与网络安全能力深度融合，发布天问系列产品，实现 AI 全面赋能。

天问是天融信所有 AI 能力的统称，以大模型、小模型、机器学习等 AI 技术能力为核心，覆盖基座大模型平台、安全软件平台、安全硬件平台、安全管理平台、天问算力平台以及相关产品应用的企业级 AI 安全能力体系。天问由天问基座大模型、天问大模型系统、云上小天、产品小天、天问智算云平台等产品组成。其中小天是所有问答类模块、系统的统称，包括云上小天、产品小天。

天问基座大模型是面向网安垂直领域的基础能力大模型，采用多年积累的数据进行训练和微调，形成安全对话和安全代码大模型，应用于安全运营、知识问答、算力管理等场

景。天问大模型系统是 AI 安全大脑，提供自动化研判、小天人机对话等功能，同时，面向天融信大数据分析系统、脆弱性扫描与管理系统、数据库审计与防护系统等产品提供大模型应用调用服务。

云上小天是依托天问基座大模型与天问大模型系统，面向天融信及各行业客户推出的安全问答与能力订阅，提供天融信产品、技术、解决方案、行业知识等服务。产品小天是依托天问基座大模型，面向天融信安全产品推出的天问助手，以组件的形式嵌入到各类网络安全产品中，以智能问答的形式，提供告警分析、处置建议、报告生成等功能。

天问智算云平台是面向大模型应用场景推出的一款天问算力系统与管理平台，可实现大模型产品私有化部署，为客户提供“硬件+软件+服务”的一体化解决方案，以一体机的形式交付，提供基础算力，并内置多种主流大模型、算力调度管理、安全网元等能力。

天融信天问大模型可为用户提供威胁检测、安全运营、知识问答、算力管理等能力。在威胁检测方面，天融信将 AI 检测引擎融入到防火墙、僵木蠕、IDS、IPS 等产品中，在各种环境中，实现对隐蔽攻击、未知威胁的精准、高效捕获和阻断。在安全运营方面，天融信天问大模型系统基于 AI 技术建立机器学习模型，让安全人员以自然语言方式获取信息，简化安全人员操作过程，提升安全日志研判效率，实现更加精准和便捷化的安全运营。在知识问答方面，云上小天可提供情报信息解读，协助客户对安全威胁进行分析、溯源，检测资产失陷情况。产品小天可智能分析客户提问，理解用户意图，快速响应客户需求，自动输出图表、文本等形式的分析结果。在算力管理方面，天问智算云平台专为大模型私有化部署场景设计的算力平台，可提供强大的计算能力和高效的模型管理能力。

15. 云起无垠

云起无垠是一家初创的从事 Fuzzing 产品开发的公司，专注于应用大模型探索安全缺陷的智能检测及修复技术，打造最懂安全的 AI 智能体。公司依托自主研发与训练的“云起 AI 安全大脑”，推出了无极 AI 安全智能体平台，涵盖安全知识问答、智能模糊测试、代码分析与生成、漏洞威胁情报等功能，以帮助企业自动化完成各类安全任务。

无极 AI 安全智能体平台融合了 AI 技术与安全智能体的理念，打造了一个集成安全知识和安全工具的智能化平台。该平台为用户提供了一个全面的安全技术支持和解决方案库，旨在帮助用户应对网络安全威胁、代码漏洞风险以及安全工具治理过程中的复杂性和挑战。

无极 AI 安全智能体平台通过其独特的智能体编排集成技术，为用户带来便利和效率。它集成了广泛的安全知识库，涵盖了最新的安全威胁情报和防护策略，使用户能够及时了解并应对各种安全风险。在代码安全方面，平台利用先进的 AI 技术自动检测和修复代码中的漏洞，减少人为审查的负担，提高代码质量和系统安全性。此外，平台还整合了各种安全工具，通过统一的管理界面简化了工具的使用流程，降低了学习成本和操作复杂性。这种集成化的安全工具治理方式，使用户能够更加有效地进行安全管理，提升整体防护能力。

16. 中国电信

中国电信的网络安全业务底子比较好，后成立天翼安全子公司，专门从事网络安全业务，在安全大模型方面也有布局，借助中国电信的安全数据、威胁情报的优势，以及相对丰富的算力资源基础，以安全运营辅助为切入点，开发了“见微”安全大模型。

目前在大模型赋能安全方面的具有应用有以下几个方面：

告警降噪研判：目前已经在电信集团和各省级指挥调度中心的 SOC 平台上应用了大模型安全降噪能力，主要是对不同设备的各类告警日志，通过大模型的能力进行判断并且生成处置建议，当前告警的压降可以达到 99%以上（100 条告警压至 1 条以下）。此外，这部分能力也是和电信态势感知系统紧密绑定的，可分为 SaaS 和私有化部署多种模式。

安全智能助手：通过问答的形式辅助安全运营，比如使用安全助手对某个单一的事件进行详细分析，也可以支持安全报告自动生成等，这部分能力可与态势感知系统集成，也可以作为独立的安全助手来应用，还可以支持 API 调用。

此外，在资产脆弱性管理等方面也应用到了安全大模型相关技术。同时，中国电信自研了星辰 AI 大模型并全面开源。

五、 企业安全大模型能力评估

（一）评估纬度

1. 安全能力

安全能力是安全大模型的核心能力，一个安全能力不够强的公司不可能训练出具有很强安全能力的安全大模型，而安全能力需要专业的人员、长期的积累以及所谓安全文化。

2. 深度学习技术能力

鉴于当前网络安全大模型往往是大模型与小模型相结合的解决方案，我们需要考虑企业在深度学习技术方面的积累。

3. 基础大模型能力

构建基础大模型的能力。

具有此类能力的厂商极为有限，对厂商的技术能力、数据获取能力、自有算力/租用算力能力均有很高要求，此能力对安全大模型来说不是必须具备的能力，但具有此能力的供应商因为对大模型的原理有更深刻的了解，在大模型的预训练与精调工作中容易获得一些优势。

4. 安全数据能力

这是构建安全大模型的核心能力之一，用于安全大模型预训练、精调的数据量与数据质量都是会影响安全大模型能力的最重要的能力之一。

5. 大模型精调能力

是构建安全大模型的核心能力之一，安全大模型往往是在基座大模型基础之上直接进行精调，甚至没有预训练过程，精调中有很多技巧，精调的过程直接关系到安全的大模型的效果。

6. 算力能力

安全大模型的预训练与精调虽然不如基座大模型训练过程需要那么多算力，但至少还是会需要数百张显卡规模的算力集群，算力能力直接关系到大模型的迭代速度。

7. 产品化能力

在中国，安全大模型更多还是以私有化部署形式存在，产品化能力是关系到产品交付给客户之后的服务成本的重要因素。

8. 用户场景的覆盖能力

用户场景的覆盖能力是一个综合技术、产品与市场能力的综合能力，要具备能解决某种安全场景的技术能力、产品能力，还要具备能把产品卖给客户的能力。

（二）国内部分网络安全公司安全大模型能力评估

1. 360 数字安全集团

2. 安恒信息

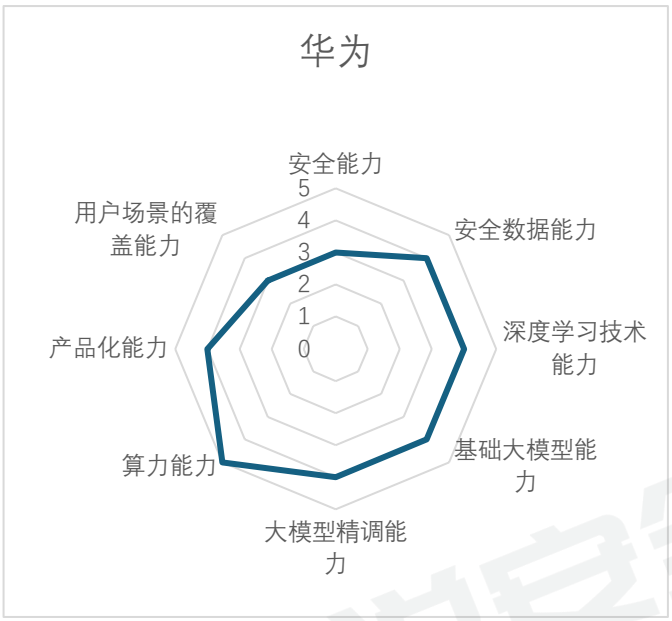
3. 海云安


海云安
安全能力
5
4
3
2
1
0
安全数据能力
深度学习技术能力
基础大模型能力
大模型精调能力
算力能力
产品化能力
用户场景的覆盖能力

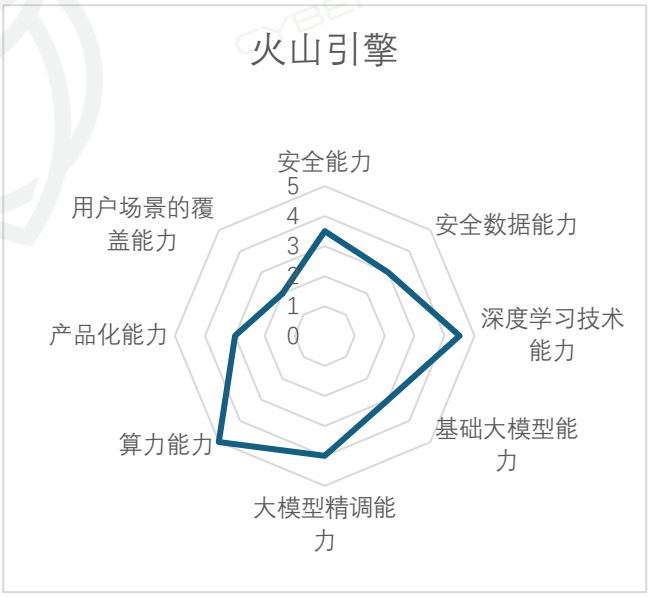
4. 华清未央


华清未央
安全能力
5
4
3
2
1
0
安全数据能力
深度学习技术能力
基础大模型能力
大模型精调能力
算力能力
产品化能力
用户场景的覆盖能力

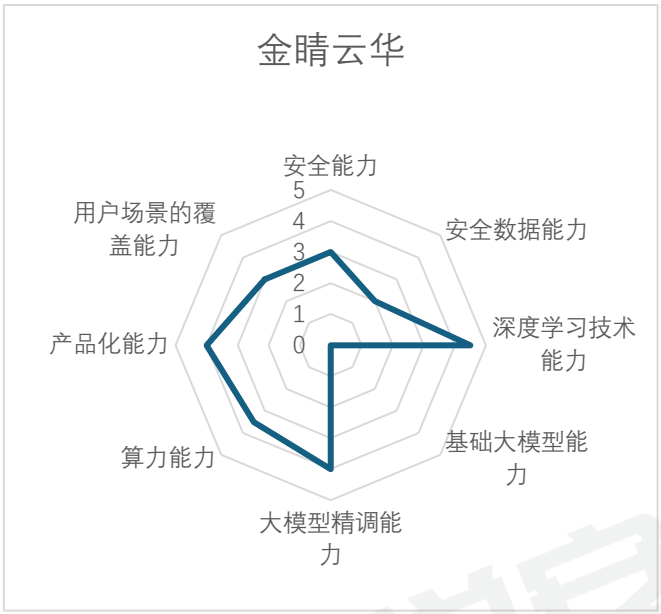
5. 华为


华为
安全能力
安全数据能力
深度学习技术能力
基础大模型能力
大模型精调能力
算力能力
产品化能力
用户场景的覆盖能力
5
4
3
2
1
0

6. 火山引擎


火山引擎
安全能力
安全数据能力
深度学习技术能力
基础大模型能力
大模型精调能力
算力能力
产品化能力
用户场景的覆盖能力
5
4
3
2
1
0

7. 金睛云华

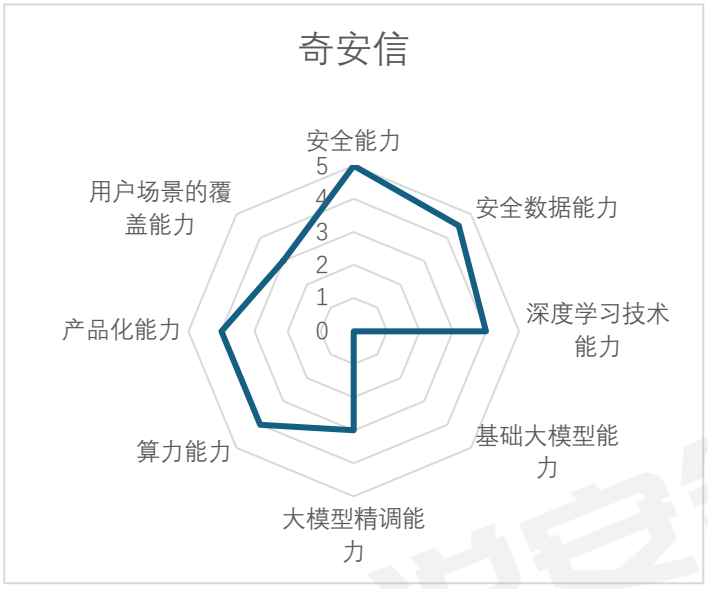


8. 酷德啄木鸟

9. 绿盟科技

10. 灵云数科

11. 奇安信


奇安信
安全能力
安全数据能力
深度学习技术能力
基础大模型能力
大模型精调能力
算力能力
产品化能力
用户场景的覆盖能力
5
4
3
2
1
0

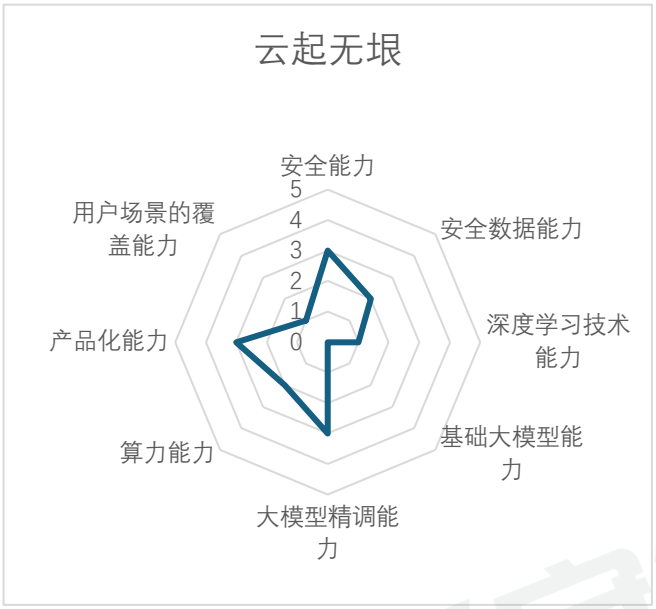
12. 深信服


深信服
安全能力
安全数据能力
深度学习技术能力
基础大模型能力
大模型精调能力
算力能力
产品化能力
用户场景的覆盖能力
5
4
3
2
1
0

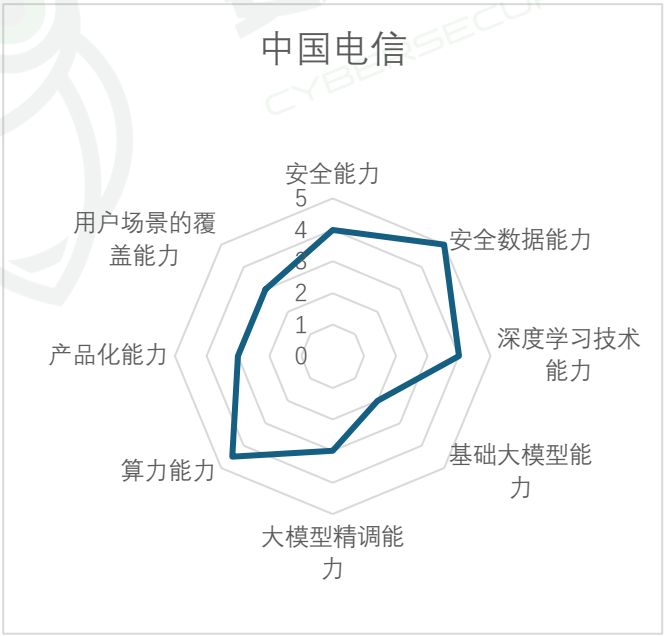
13. 天融信

14. 腾讯安全

15. 云起无垠



16. 中国电信



(三)国内安全大模型产品推荐供应商

1.安全运营大模型推荐供应商

供应商为：深信服、安恒信息、金睛云华、奇安信、天融信、绿盟科技、中国电信、华为、360

2.威胁检测大模型推荐供应商

供应商为：金睛云华、深信服、安恒信息、360、奇安信、天融信

3.数据安全大模型推荐供应商

供应商为：安恒信息、腾讯安全、火山引擎

4. 邮件安全大模型推荐供应商

供应商为：灵云数科

5. 自动渗透大模型推荐供应商

供应商为：绿盟科技

6. 漏洞挖掘大模型推荐供应商

供应商为：华清未央、云起无垠

7. 安全开发大模型推荐供应商

供应商为：海云安、华清未央、酷德啄木鸟、云起无垠

潜在供应商（已内部使用，待产品化）为：奇安信、安恒信息

六、 解决方案与案例

(一)360 安全大模型

1. 应用场景

360 安全大模型定位于：

- 赋能产品、平台提升看见的能力。
- 为安全运营提供高度自动化的处置能力。
- 总结知识，加持于人，提升组织贡献，解决安全专家人力不足水平不足的问题。

围绕客户的网络安全问题，主要应用再如下几类场景：

- 安全问答：

安全基础知识解答、产品使用方法解答，提供 IP、域名、网址、样本、漏洞、证书等情报、法规、政策解读、辅助运营工程师分析研判，提高人效。

- 威胁检测：

告警研判降噪、高级威胁发现和溯源、恶意邮件检测、恶意流量检测。

- 安全运营：

告警解读、安全事件处理、处置操作执行、资产运维管理、报告编写、安全处置建议、预案处置能力、智能化企业安全运营，自动监测威胁、追溯，并最终提供解决方案。

其他场景还包括代码安全、模型安全等。

2. 技术方案

360 安全大模型产品是以自己研发的多专家协同(CoE)架构模型为中枢，通过检索增强生成（RAG）和工具增强生成（TAG）,联动 360 安全知识库和工具库，通过智能体框架连接、配置、驱动、协同各类安全工具产品解决企业网络安全问题。

3. 部署形态

360 安全大模型产品以软件形态私有化部署在用户内网。包括 CoE 架构模型、智能体应用。

在本地化构建以安全大模型为核心的智能体框架，同时部署相应的 XDR 分析平台，配套 EDR 和 NDR 探针等工具，充分发挥智能体 RAG、TAG 的各种能力，实现像拥有高级专家群一样的安全能力。

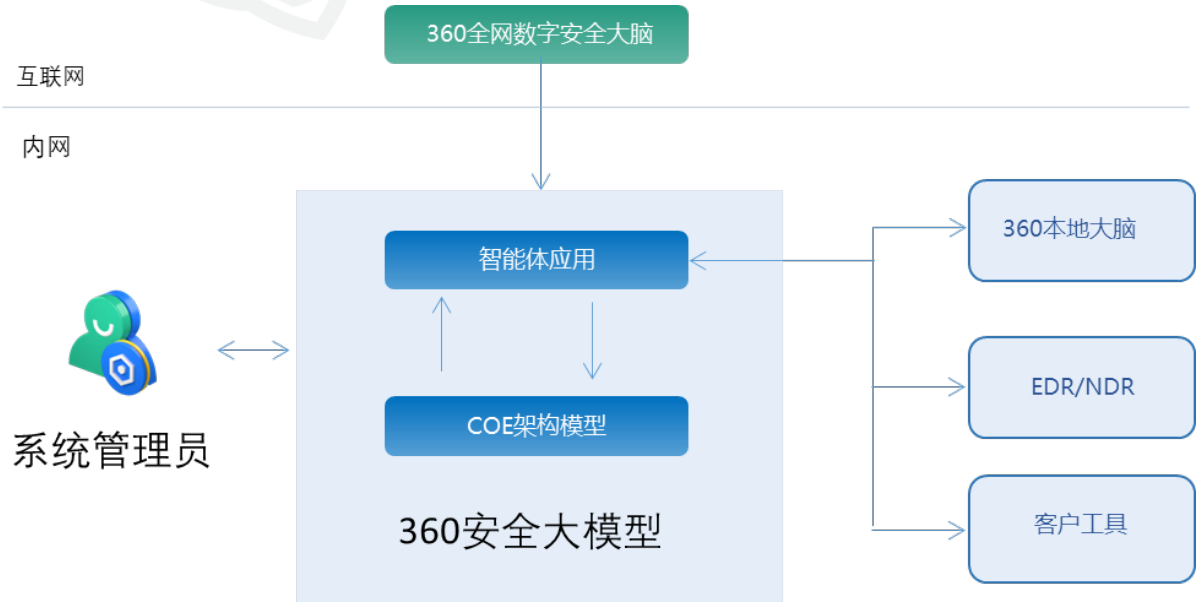


图 14 360 安全大模型内外网联动示意图

- 赋能产品

360 安全大模型赋能“360 本地安全大脑、终端安全产品、流量检测产品、网络空间测绘、资产漏洞管理平台”等不同品类产品，同时也对 360 远程托管运营服务进行了升级赋能。

- 独立应用

360 安全大模型除了对 360 安全产品进行赋能以外，也提供例如问答助手、辅助研判工具使用。

4. 硬件要求

可以应用 A6000、L20 这类消费级显卡，最小的硬件配置要求 CPU80C、GPU48G*4、内存 256G、硬盘 SSD512G，根据不同任务类型的处理量调整 GPU 资源；支持国产化 GPU，例如升腾 910。

5. 效果评估

安全大模型在威胁检测方面，无论是恶意流量、钓鱼邮件还是终端告警等方面的 F1-Score 都在 95%以上；

安全大模型在安全运营方面，人工平均响应效率提升 100%，事件平均检测效率和事件自动化响应率提升 200%；

6. 特色

- 独创 CoE 类脑框架模型
- 拥有最全面的高质量安全大数据

- 可实现类安全专家完成自动化高级威胁狩猎
- 最丰富的 AI 安全应用经验

7. 标杆客户

金融行业作为技术先锋行业，XX 客户属于金融行业安全走的靠前的企业，在新技术研究和应用上一直保持领先。随着大型模型技术的兴起，黑客们也开始利用 AI 工具提高攻击效率，使得传统安全问题变得更加紧迫。这种新形势要求企业加强对安全领域的关注和投入，及时应对并应用最新技术来保护金融系统的安全。

痛点分析：

安全运营是安全领域里面自动化程度最低，投入精力最多的工作类型，但安全效果难保证，难有工作成果产出。企业当前在安全能力建设方面存在如下几个短板：

- 安全能力不足，高级威胁日益增多，现有设备和人员安全能力不足。
- 运营效率低下，安全运营基本依赖人力，无法实现自动化，效率低下。
- 安全专家缺乏，安全人才培养周期长，高端安全专家稀缺。

解决方案：

从企业当前面临的安全能力和运营效率问题切入，综合“数据、场景、模型、智能体”四个方面核心要素，就企业当前面临的安全能力不足、运营效率低下、安全专家缺乏等痛点问题进行系统分析，最终形成以安全大模型为“大脑”，构建智能体框架，通过智能体框架的任务编排、指令调度、记忆存储等能力，调用安全知识、工具，充分发挥检索增强（RAG）、工具增强（TAG）的各种能力，对安全大模型的结果输出进行纠错和能力增强，实现强大的安全专家能力。

图 15 360 安全大模型架构图

8. 客户价值

打造高端安全能力，提升“看见”威胁的能力

360 安全大模型和其他安全产品联动，利用内部各类安全检测专家分区，对安全产品的告警、日志数据进行分析以发现潜在的威胁和攻击行为。对告警或者高危安全事件执行威胁猎杀任务，发现真实攻击意图，帮助产品发现过去难以发现的攻击行为。为企业配备数字安全专家，提升“看见”威胁的能力。

重塑智能安全运营，提升安全运营效率

通过 360 安全大模型，对日常的网络安全运营流程提供辅助支持。通过问答方式帮助管理员迅速完成重要任务，同时通过告警解读、事件研判、HQL 语句自动生成等功能，帮助管理员更准确更快速的对告警进行识别，提升日常运营效率。

总结全面安全知识，降低技术门槛

360 安全大模型集成了 360 多年的安全大数据、威胁情报、攻防对抗知识等专业知识以及安全运营经验、HW&重保经验、应急预案等行业经验。针对代码能够实现漏洞检

测并给出修复建议，提高企业内部代码编写安全水平。让管理员可以快速、便利的获取到所需相关安全知识，提高安全运营人员水平，解决安全专家人力不足水平不足问题。

(二)安恒恒脑安全大模型介绍

1. 应用场景

安全运营，包括日常告警研判、报文检测、事件调查、联动处置、数据分类分级和报告撰写等场景。

2. 技术方案

恒脑大模型是基于通用大模型经重新训练、精调等方式形成的安全领域垂域模型，基座模型通义千问 7B 和 72B，针对不同的应用场景采用不同的模型，如智能问答、报告撰写等采用 72B 模型，如报文检测使用 7B 模型。

异构多微调模型的统一推理加速，通过构建包括全参数模型、多个 PEFT（Parameter Efficient Fine Tuning）模型（如 LoRA 模型、Prefix-Tuning、Adapter-Tuning）在内的混合模型，执行统一批次处理（batching）实现全局推理加速，构建基于多模型协调组织的 GPU 算力优化方案，实现数倍的 GPU 算力提升。

图 16 安恒恒脑 2.0 示意图

3. 部署形态

恒脑安全垂域大模型系统部署形态：纯软、软硬一体两种模式，目前已与平台级产品 Ailpha 政企/监管态势感知和 SOAR 平台深度融合，以插件形式与安恒 30 多款产品完成对接。

4. 硬件要求

算力支持情况如下：

1）支持英伟达 A100、A800、H100、H800、RTX4090、L20 等 GPU 算力，恒脑一体机出货采用标配 L20*6 卡（支持扩展到 10 卡）。2）支持英特尔至强可扩展处理器 CPU 应用于 14B 规模以下参数模型。

图 17 安恒恒脑对 INTEL 处理器的支持

3）支持华为昇腾 310 和 910 系列算力。

图 18 安恒恒脑对国产算力芯片的支持

5. 效果评估

恒脑大模型在告警研判场景，准确度达到 98%，性能相较人工有 100 倍提升；在数据分类分级场景，准确度达到 90%，性能相较人工有 30 倍提升。

6. 特色

相较其他竞品，安恒信息通过构建 MoE 架构的安全大模型，可基于业务场景灵活选用各种参数规模的模型，通过开发智能体中台，支持各类安全工具和安全知识的灵活接入。具备多安全 Agent 协作能力，任务完成度达 100%。安全能力理解调度强，可调度安全能

力不少于 100 种，可理解安全知识和数据不少于 300 种。覆盖场景全，重保场景覆盖度超 80%，日常安全运营场景超 90%。

7. 标杆客户

- 杭州亚运会 ITCC 网络安保指挥中心，杭州亚运会期间分析超 800 亿条安全日志、监测拦截超 2600 万次网络攻击、响应专家 34792 次提问，辅助处理 287 起安全事件，综合提升安保小组 57%的工作效率。
- 浙江移动 IT 中心告警研判，通过使用告警研判 AI Agent 基于现场实测，准确率达到 98%等同高级安全专家水平，性能 120 万条/天（3 张英伟达 A100 显卡）是人工（20 人团队）的 100 倍
- 某金融机构数据分类分级，采用数据分类分级 AI Agent，10W 字段实测，项目实施周期从业界平均超 100 人天下降至不到 20 人天，识别准确率提升至 90%。

（三）金睛云华安全运营智能体案例

1. 应用场景

金睛云华的安全大模型共分两个，检测大模型和运营大模型，检测大模型主要用于各种检测场景：

- 流量检测：可以针对恶意软件流量，攻击流量，特种应用流量（如 VPN）等进行检测，可以直接检测加密流量，不需要解密。
- 脚本检测：可以检测各种加密、变形的攻击脚本，包括 javascrpit,sql 注入等。
- 二进制检测：能够直接针对病毒类二进制文件做检测。

运营大模型主要应用于以下场景：

- **告警降噪：**通过智能化分析和处理，大幅减少安全告警中的误报，提升告警的有效性和精确度。
- **安全运营：**用于自动化处理和分析安全事件，辅助安全运营人员进行快速响应和决策，显著提高运营效率。

2. 技术方案

检测模型采用 BERT 技术，以程序语言为基础进行预训练，并使用流量 pcap,脚本，二进制文件等标签数据做 SFT,形成基本模型，当前该模型完全自主训练，非基于开源。

运营模型基于国外和国内的开源基础模型，进行了二次预训练和微调，技术路线如下：

- **基座模型：**采用 Llama、Mistral、Qwen 等先进大模型作为基座，确保基础模型的高效性和可靠性。
- **二次预训练：**在通用大模型的基础上，使用大量安全相关数据进行二次预训练，使模型具备更强的安全领域知识。
- **微调：**针对告警降噪等具体场景，进行精细化的微调，优化模型在实际应用中的表现。

在模型的基础上，进行了大量的外围开发，形成安全运营的智能体，能够对接客户已有的安全设备，本地数据，情报等，具备智能分析和研判能力，可以直接调用工具，或者对安全设备的配置优化等提出建议（当前阶段暂不支持自动化配置，但能力已经具备）。

图 19 金睛云华「安心 CyberGPT」大模型

图 20 金睛云华「安心 CyberGPT」大模型

3. 部署形态

金睛云华的安全大模型支持本地化部署，形态灵活多样：

- 纯软件部署：如果客户已有 AI 服务器，模型可以作为独立的软件系统部署，便于企业内部使用。
- 软硬一体：提供软硬件一体的解决方案，配备专门的具备 GPU 的设备，以提升整体处理性能。
- 系统整合：可与现有的安全信息和事件管理系统（SIEM）等安全系统进行无缝整合，增强整体安全防护能力。

4. 硬件要求

为了支持金睛云华的安全大模型高效运行，硬件要求如下：

- **GPU 支持：**基于 NVIDIA 4090 GPU 即可运行，具备良好的性价比和计算能力。
- **国产算力支持：**支持如华为昇腾（Ascend）等国产 AI 加速卡，以满足数据安全及自主可控的需求。

5. 效果评估

金睛云华的安全大模型在实际应用中取得了显著的效果：

- **检测能力明显提升，**尤其是针对加密流量及 Web 攻击的检测水平明显上升。
- **告警降噪效果：**有效减少了 95%以上的误报，显著提升告警的准确性。
- **安全运营提升：**人效比可以达到 1:10,即一台机器可处理相当于 10 人的工作量，大幅提升安全运营效率，缩短事件研判和响应时间。

6. 特色

金睛云华的安全大模型在以下方面具有独特的优势和亮点：

- 完整的检测能力：针对流量，脚本，二进制等的多模态检测能力。准确率更高，误报更低。
- 先进模型技术：基于 Llama、Mistral、Qwen 等最新大模型，具备更强的理解和分析能力。
- 高效预训练与微调：结合安全领域的特定需求，进行二次预训练和精细微调，确保模型在实际应用中的高效表现。

- 灵活部署：支持多种部署形态，特别是本地化部署，能够有效解决数据安全的问题，满足不同企业的使用需求。
- 国产化支持：全面支持国产算力平台，保障数据安全及自主可控。

7. 标杆客户

金睛云华的安全大模型已在多个标杆客户中取得成功：

- **某大型银行：**通过使用金睛云华的安全大模型，显著减少了安全告警中的误报，提高了安全事件的响应速度，保障了银行系统的安全稳定运行。
- **某政府部门：**在部署金睛云华的安全大模型后，安全运营效率大幅提升，成功应对多次网络攻击，确保了政府数据和系统的安全。

金睛云华的安全大模型凭借其卓越的技术和显著的应用效果，赢得了客户的高度认可和信赖。

(四)深信服安全 GPT

1. 应用场景

深信服安全 GPT 基于海量网络安全知识和最佳实践流程的学习，构建安全行业的智能研判和调度中枢，可对接检测响应、态势感知、端点安全、边界安全等系统，如可扩展检测响应平台 XDR、安全感知管理平台 SIP、下一代防火墙 AF、统一端点安全管理系统 aES、安全托管服务 MSS 等安全产品与服务。

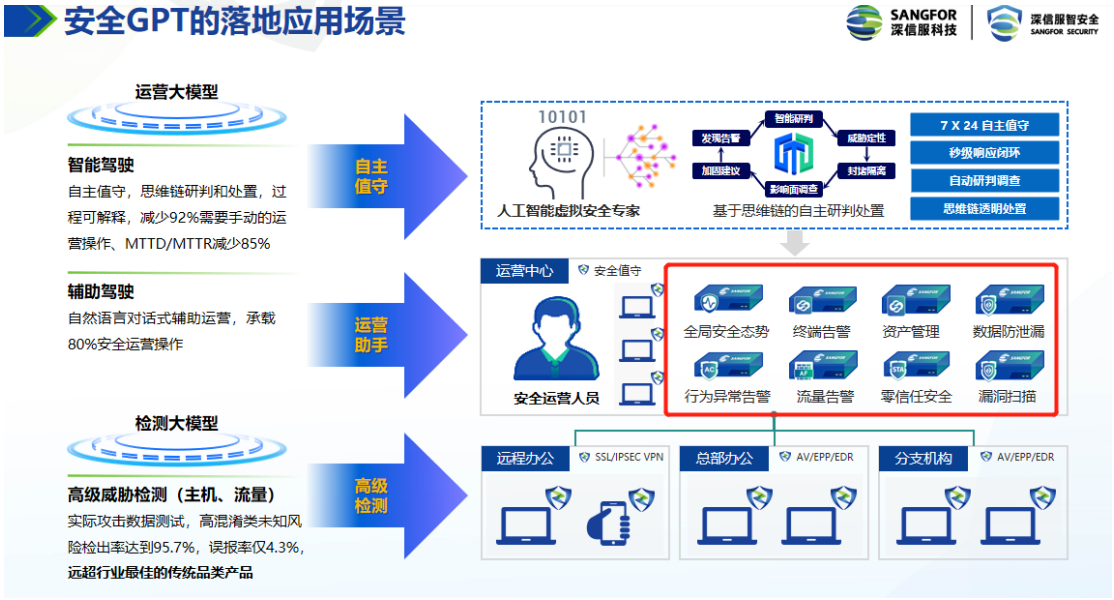


图 21 深信服安全 GPT 的落地场景

- 未知威胁检测。通过海量 HTTP 流量、日志等数据预训练和微调，安全 GPT 可以检测流量中的高对抗攻击行为，如 Web shell 混淆、编码混淆、XSS 混淆攻击等。
- 高对抗钓鱼检测。安全 GPT 支持有效识别邮件、文件、网页等多种途径钓鱼风险，基于端侧采集的文件落盘及上下文行为数据，从敏感信息、意图、写作风格等方面进行钓鱼意图检测，实现整个攻击过程的自动或半自动遏制及响应闭环。

- 安全运营对话助手。通过自然语言的方式提供对选定事件的分析、研判，大幅加快事件处置效率。
- 7*24 小时自主值守运营。大模型通过理解网络行为的上下文环境，全面自主值守支持 7x24 小时自主值守、可实现秒级响应闭环、支持自动研判调查、思维链透明处置。

深信服安全 GPT 有两种工作模式，第一种是对话式的“辅助驾驶”模式：

图 22 深信服安全 GPT 的“辅助驾驶模式”

第二种是基于自主研判思维链的“智能驾驶”模式：

图 23 深信服安全 GPT 的“智能驾驶模式”

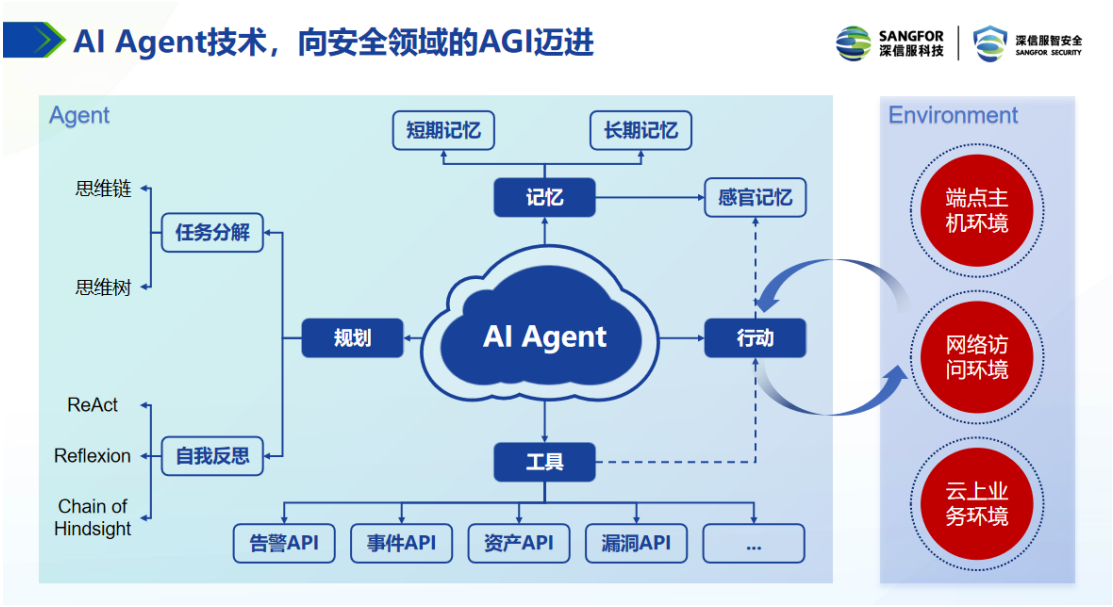


图 24 深信服安全 GPT 的 AI 智能体

传统人工 VS 辅助驾驶 VS 智能驾驶安全运营效率对比如下：

图 25 两种模式的安全运营效率对比

2. 技术方案

基于多个基座模型的定向训练形成辅助运营、流量解读、钓鱼意图分析等多个专家模型，安全 GPT 是围绕多个基座模型的连接和组织形成安全智能体，通过 APO 算法实现了 Prompt 自动化优化，减少了人工干预的需求，并全面采用模型加密和模型水印技术确保了模型的安全性。

深信服安全 GPT 通过 XDR 平台进行第三方适配，基于 Open XDR 技术可极大提升对接三方组件与对接自有原生组件安全效果的一致性，可充分利旧已有各类组件，保护投资，且具备数据质量评估的能力，最大程度发挥现有安全产品/数据的价值。

3. 部署形态

深信服安全 GPT 支持 SaaS 和本地（软硬一体）部署，根据不同分支大模型的能力不同，可接入当前各类安全系统（如 XDR、SIP、aES 等），完成流量威胁检测、主机钓鱼检测、对话式安全运营、自动值守大屏等功能。

4. 硬件要求

安全 GPT 大模型仅需数张 4090 消费级显卡即可本地部署运行。国产化算力已全面适配天数智芯 MRV100、华为昇腾 300i DUO 等型号。

在 8 张 4090 级别显卡支持每天 3 万告警处理。

5. 效果评估

流量检测方面实际攻击数据测试，精准率达到 95.7%，误报率仅 4.3%，远超行业最佳的传统品类产品，隐蔽威胁不漏过。主机检测方面，对没有恶意特征的邮件与高免杀对抗的文件场景，检出效果大幅提升。三万高对抗钓鱼样本测试中，检出率达到 94.8%，误报率小于 0.1%，正报样本准确率是传统防钓鱼类产品的 4 倍多。

安全运营过程利用安全 GPT，脱离高重复性的工作，聚焦高价值的创新，减少 92%需要多次手动的运营工作、MTTD/MTTR 减少 85%。依靠于人工实现安全事件全流程闭环约需要 3 小时，通过安全 GPT 的辅助驾驶可缩短至 5-10 分钟，若进一步启用安全 GPT 智能驾驶功能则可控制在 30 秒。

6. 特色

深信服安全 GPT 是国内通过网信办双备案的安全大模型，分别是《生成式人工智能服务管理暂行办法》备案、《互联网信息服务深度合成管理规定》备案。深信服大模型的训练及优化过程基于自研的 AIGC 高性能计算平台，综合训练成本和训练周期相比于传统优化方法节省 50%。2023 年即完成标准化的产品交付，截至 2024 年 6 月累计上线客户 150+家。

7. 标杆客户

某国家部委利用深信服安全 GPT 结合安全运营平台，构建智能调度控制中枢，支持对日均超 1 亿条的安全原始日志数据的分析总结，并对增量数据分析推理。使用过程中，检测大模型试运行测试精准率>95%，误报率<4%，独报告警占比达 82.8%，并在实际业务环境中发现高混淆攻击案例。安全 GPT 自主研判，实现告警降噪 99.8%以上。高级安全运营人员精力充分释放，工耗降低 25%，安全 GPT 补充了夜间安全监测研判处置力量，实现全天候 7*24 小时安全值守。

某顶尖制造企业通过自然语言对话与分析系统交互，快速获取安全数据、识别威胁模式，从而大幅提高安全运营团队分析研判的效率。安全负责人表示，深信服安全 GPT 让运营人员在广度和深度上都能做全局把控。在广度上，少量运营人员即可守护数万资产，每天只需关注安全 GPT 逐一研判后定位的日均 100 条高危告警，准确度超过 97%。在深度上，安全 GPT 对任意一条告警都可解释，直观呈现完整分析过程，帮助运营人员更好理解攻击意图、完成研判决策。

(五)天融信天问安全大模型方案

1. 应用场景

天问是天融信所有 AI 能力的统称，以大模型、小模型、机器学习等 AI 技术能力为核心，可为用户提供威胁检测、安全运营、知识问答、算力管理等能力。

1） 威胁检测

天融信防火墙、僵木蠕、IPS、IDS 等产品内置 AI 安全威胁检测引擎，进行关联分析、识别 DGA 域名、隐蔽通道和恶意加密流量，阻截攻击者的入侵行为，利用 DGA 识别模型对域名进行精准的检测判断，识别出 DGA 域名，完成对 C&C 通信过程的阻断，从而实现对高级威胁的防御。

2） 安全运营

天融信天问大模型系统提供自动化研判、小天人机对话等功能，大幅提升安全日志研判效率，实现更加精准和便捷化的安全运营。针对海量安全日志，系统基于 AI 技术建立机器学习模型，学习历史告警研判结果与告警来源、告警规则、攻击源 IP 之间的关联关系，建立来源、规则、攻击源 IP 可信性，以此来自动评估新告警的可信度，将高可信度告警推荐给安全人员进行人工确认，简化安全人员操作过程，加快研判信息获取。

3） 知识问答

云上小天是面向天融信及各行业客户推出的安全问答与能力订阅服务，依托天问基座大模型能力，天融信将自然语言理解、NER 技术应用于威胁情报的分析、生产，大幅提升威胁情报的精准度和产出效率。产品小天是基于天问基座大模型，以组件的形式嵌入到各类网络安全产品中，可智能分析客户提问，理解用户意图，快速响应客户需求。

4） 算力管理

天融信天问智算云平台专为大模型私有化部署场景设计，可提供强大的计算能力和高效的模型管理能力，平台能够管理异构算力 GPU 资源池，并内置开源主流大模型，可帮助用户简化大模型准备过程。通过该平台，用户能够快速、高效地构建和部署 AI 应用，进行 AI 训练和 AI 推理。

2. 技术方案

天融信下一代可信网络安全架构（NGTNA）是以网络安全为核心、大数据为基础、云服务为交付模式，为客户构建具备全面感知、智能协同、动态防护、聚力赋能的可信网络安全保障体系，该架构可覆盖物理环境、云环境等应用场景，为客户业务系统的安全、可持续性运行实现赋能。

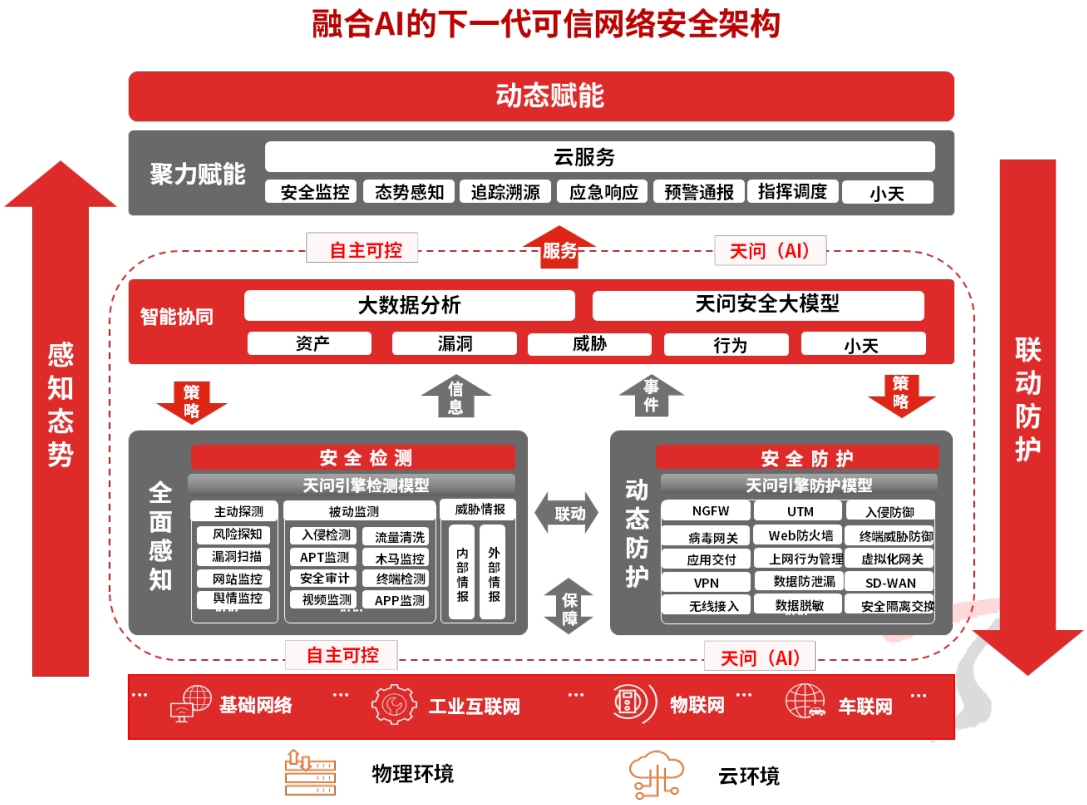


图 26 天融信融合 AI 的下一代网络安全架构

天融信 NGTNA2.0 架构融合 AI 关键技术、平台及产品能力后，在感知、防护、分析、联动等层面能力全新升级，实现 AI 全面赋能，进而带来体系架构整体能力的提升。详细 NGTNA2.0 架构 AI 能力升级如下：

架构层	AI 联动能力
聚力赋能	新增云上小天，以云服务的形式为天融信及客户提供安全问答与能力订阅。
智能协同	新增天问大模型系统，提供大模型调用服务。
	大数据分析系统 AI 能力升级，向下驱动各类 AI 安全引擎，同时支持对各类探针数据进行深度分析。
全面感知、动态防护	防火墙、IPS、IDS、僵木蠕等安全产品内嵌基于机器学习的 AI 检测引擎，满足本地化的安全防护需求。
	产品内嵌产品小天，以 AI 助手提供本地的知识问答服务，提升客户的安全运营效率，同时与天问大模型系统、大数据分析系统实现联动，调用大模型知识问答服务。

3. 部署形态

天问系列产品支持本地部署、云端订阅、嵌入式部署，威胁检测引擎本地部署在防火墙、僵木蠕、IPS、IDS 等设备中，天问大模型系统、云上小天以云服务方式部署，提供云端订阅服务，产品小天嵌入到各类网络安全产品中，以组件形式为客户提供智能问答服务。

4. 硬件要求

支持 Intel CPU +NVIDIA GPU、鲲鹏 CPU+昇腾 GPU,模型推理最低 GPU 显存 16GB 及以上。

5. 效果评估

天融信天问大模型结合机器学习、深度学习等技术，利用大规模数据进行训练和优化，以提升安全研判分析的效果和能力，在实际应用中带来多方面的效果：

1. 智能自动可以更准确地识别告警，从而减少误报和漏报的情况，提高真实告警的检测效率。实际使用显示，每人天研判告警数目可提升一倍以上。
2. 通过学习大规模的数据，可以识别和理解更复杂、更隐蔽的安全威胁和异常行为，从而提高研判分析的准确性和精度。实际使用显示，有效告警率可提升 5 倍以上。
3. 可以识别潜在的安全威胁和漏洞，帮助安全团队及时发现并解决安全问题，降低安全风险。

6. 特色

- 采用于机器学习、深度学习、自然语言处理、大语言模型、小语言模型、超微语言模型等多种 AI 技术。
- 将 AI 技术全面应用于天融信基座大模型平台、安全软件平台、安全硬件平台、安全管理平台、天问算力平台等平台中，提升平台 AI 能力。
- 借助平台能力升级，各安全产品即可快速具备全面的 AI 安全能力。
- AI 能力覆盖全面，涵盖威胁检测、安全运营、知识问答、算力管理等场景。

7. 标杆客户

天问系列安全产品覆盖威胁检测、安全运营、知识问答、算力管理等应用场景，目前已在金融、运营商、能源等行业内多个大型客户落地应用，并取得了一定的成效。

在某金融客户场景中，用户部署了天融信大数据分析系统，通过内置天问大模型，实现大模型赋能；增加产品小天功能，实现人机对话，支撑安全问题智能问答；建立告警分析、漏洞分析、日志分析等插件，对人机对话中的安全运营指令进行接收响应。通过该系统部署率先在安全管理工作中引入了生成式人工智能技术，实现了行业内试点示范，通过项目建设提供的虚拟安全专家，提升了安全运营效率。

（六）中国电信安全见微安全大模型

1. 应用场景

电信安全见微安全大模型通过 AI 赋能安全态势感知平台，整合用户网络安全类数据，开展面向不同安全场景的多维智能分析。聚焦安全运营场景中的告警疲劳、效率低下、自动化程度低等问题，让大模型充分发挥价值，实现自动化异常行为分析、自适应防御策略生成、告警评估和攻击研判，让安全运营人员从劳动密集型的任务中得以解救，突破运营效率瓶颈，全方位赋能中国电信的网络安全保障水平。

2. 技术方案

根据不同的应用场景可选择不同的基座模型，如 Llama3-70B、Qwen1.5-72B 以及 Telechat-12B 等。在预训练和微调过程中贴合用户生产环境数据特点，训练出更符合用户使用需求、更能解决用户痛点的私有化安全大模型。

整体技术架构如下图，数据源采用运营平台各类原始日志及告警数据作为大模型输入，经数据清洗——告警研判——告警聚合——事件生成等流程后，通过 API 接口反馈给运营平台。

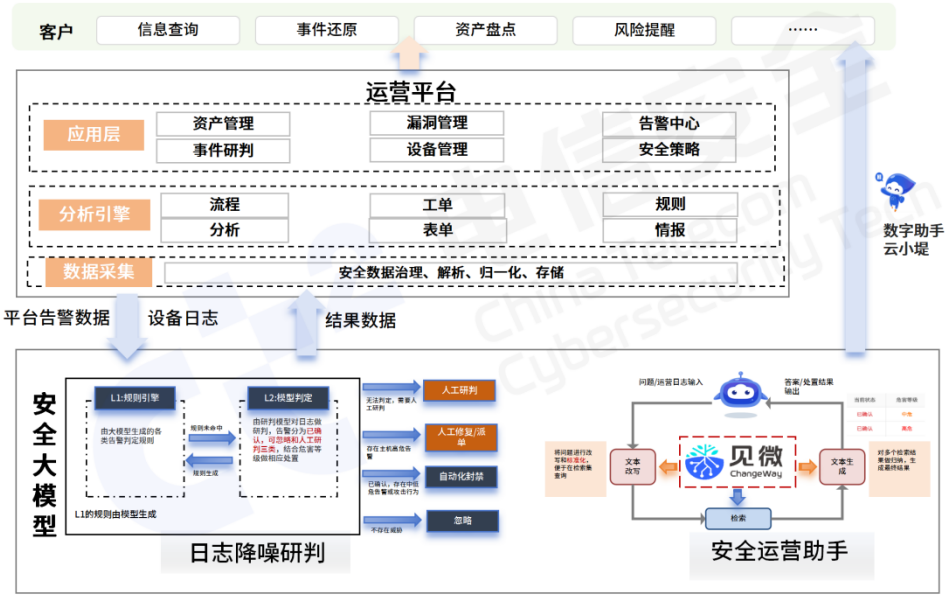


图 27 电信见微安全大模型项目技术架构

1）告警筛选

筛选的目的是从众多告警中识别出真正需要关注的告警，去除误报，避免信息过载。使用电信安全见微安全大模型后，将在筛选阶段优先去除误报，避免信息丢失。进入研判阶段，则以真实性为优先级。

2）重塑聚合

告警聚合是将多个相关联的告警合而为一，以减少告警处置工作量进而提升处理效率。聚合核心是生成告警指纹的生成算法。此阶段大模型会根据告警信息动态生成一个或一组告警指纹，后续可根据指纹优先级对告警进行归并。

3）事件研判

事件研判是对聚合后的告警信息进行深入分析，判别客体由原始告警转变为事件，进一步确定事件的真实性、优先性、危害性。

4）事件处置

处置阶段是对经研判确认需采取行动的事件进行处理。此阶段将使用电信安全见微大模型将研判结果与 Agent 处置工具结合自动化处理掉。对于无法自动化处理的告警，会生成处置建议辅助人工进行处置。

5）运营分析

运营分析阶段是对整个告警处理过程进行回顾和评估。通过对全局告警进行全面分析，从而优化未来的响应策略和相应流程。通过自动收集关键数据点和生成易于理解的报告，电信安全见微安全大模型将帮助使用者快速确认安全事件，总结事件处理的效果，并形成对应的文档记录，供使用者作进一步事件复盘。

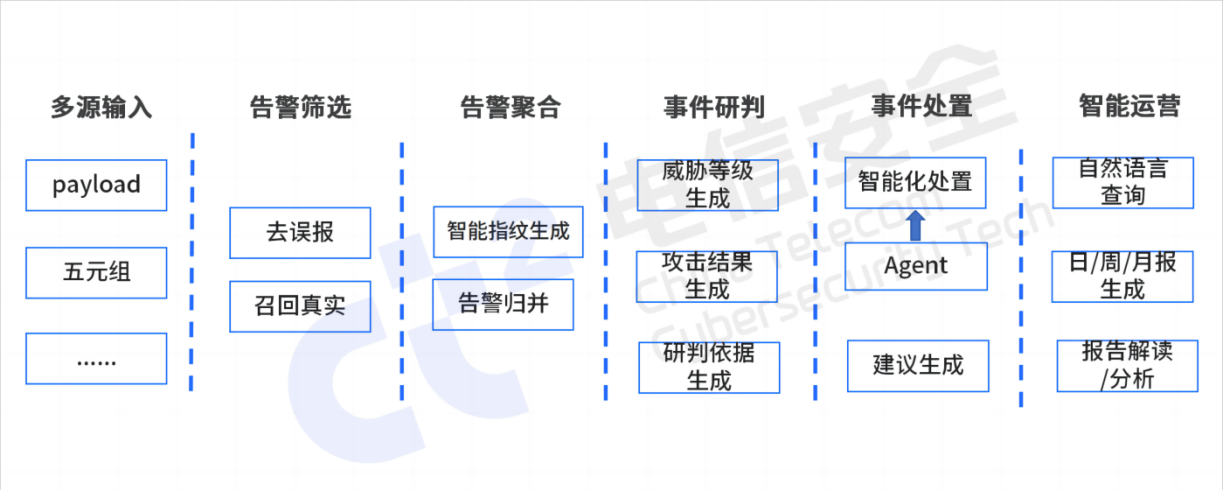


图 28 见微安全大模型侧具体路线图

3. 部署形态

电信安全见微安全大模型采用纯软件形式进行交付，与安全运营平台结合，形成从威胁识别、威胁分析、告警压降到事件生成的自动化处置能力。

4. 硬件要求

推荐：英伟达 A100 80G 4 张，或同等算力资源替代。

低配（仅推理）：V100 32G 4 张，或同等算力资源替代。

国产型号尚在适配中，具体算力需求根据业务需求调整。

5. 效果评估

在告警研判方面，利用大模型分析能力构建高置信的告警发现引擎，从原始日志或告警中筛选出真实告警，并给出研判依据和处置建议。见微安全大模型研判准确率超 98%，提供的处置建议采纳率达 50%。

在告警压降方面，通过指纹、攻击链等智能聚合方式，协助安全运营平台实现告警压降能力，降噪比可达到 99%以上，整体运营效率提升 20%以上。

6. 特色

阡陌 ChangeMore

- 电信安全见微安全大模型基于中国电信“阡陌数聚”大模型数据集，汇聚万亿级的运营商大网安全日志、威胁情报等数据，拥有区别于传统安全厂商的海量、多样化的基础数据集。
- 插件化、服务化落地模式，无需改变客户现有的安全大模型解决方案，能实现快速落地应用，支持多类平台、更具普适性。

7. 标杆客户

项目背景：

由于安全监控需要，中国电信股份有限公司四川分公司（四川电信）安全运营人员每日需处理海量由于安全设备接入而产生的原始告警日志。原有安全运营平台无法在短时间内实现对告警信息的有效处置。

项目成效：

接入见微安全大模型后，四川电信告警压降率达到 99%以上，研判准确率达到 95%以上。

免责声明

本报告所用调研数据均采用样本调研方法获得，数据分析和相关结论因受样本来源和数量的影响，未必能够完全或唯一反映真实的行业及市场现状。所以，本报告只提供给个人或单位用于必要参考，数说安全不对任何依据本报告所作的其他分析研究和判断决策负责。

致谢

本报告的数据采集工作得到了各界的大力支持，各项调查工作得以顺利进行，在各相关单位、调查支持网站以及媒体等的密切配合下，基础资源数据采集才能及时完成。在此， 谨对他们表示最衷的感谢！

数说安全介绍

专业的网络安全产业研究平台，以数据为基础，为网络安全监管部门、网络安全企业、网络安全产品与服务客户、网络安全资本市场等受众提供研究报告、顾问咨询、媒体传播、数字化营销工具等服务。

关注公众号

数说安全合作邮箱：ssaq@geniuscybertech.com

赛博英杰合作邮箱：connect@geniuscybertech.com