数世咨询
Digital World Consulting

LLM驱动数字安全

2024

AI安全系列报告

LLM-Driven Digital Security

报告编号：DWC_20240507
主笔分析师：靳慧超（战略分析师&合伙人）

报告信息

报告方向：AI 安全

报告类型：研究报告

报告名称：LLM 驱动数字安全

报告编号：DWC_20240507

主笔分析师：靳慧超（战略分析师&合伙人）

分析团队：数世咨询-数字安全研究院

智库支持：数字安全百人会

报告审核：李少鹏（首席分析师&创始人）

阅读须知

- 本报告的研究方向聚焦于如何通过 LLM（Large Language Model）来赋能我国数字安全，以及怎样应用数字安全这一垂直领域的各类 LLMs 来切实提升我国数字安全能力。关于 LLM 自身的安全性，以及如何保障各类 LLMs 训练和应用的安全并不是本报告研究的方向，相关内容敬请关注数世咨询后续 AI 安全（LLM 基础安全、AI 治理等方向）相关研究。
- 本报告分析使用的相关数据（截止日期 2024 年 4 月 7 日）从两方面获取，一是公开信息的收集，二是相关供应商的沟通。
- 据数世咨询粗略统计，现阶段国内具备 LLM 研发或应用能力的数字安全供应商有 28 家左右。本次入选报告的有 17 家，有些具备 LLM 相关应用能力的供应商并未出现在本报告中，是因为通过深入的沟通后，由于供应商经营战略以及品牌建设和产品规划的原因，暂不适合参与本次报告。
- 本报告调研的供应商包括两类，一是和数世咨询保持有效沟通的，二是通过公开信息搜寻的。如果有具备 LLM 相关的安全应用能力但并未收到数世咨询调研邀请的供应商，欢迎联系本报告主笔分析师（16601182683 微信同号）进行交流。

本报告入选标准

- 具有 AI 研究能力。
- 具有 LLM 驱动数字安全的自主研发能力。
- 投入了一定规模的资源，如算力、人力等。
- 产品具备商业化能力，已有真实落地或试用案例。
- 接受数世咨询的调研与访谈，并承诺提供数据的真实性。

安全大模型卓越能力供应商

（按品牌首字母排序）

关键发现

- ✓ 虽然“安全大模型”（LLM 驱动的安全能力）的应用还处在早期阶段，但用户方面已经展现出较强的采购意愿。这样的现状主要来源于 LLM 涌现出的新兴能力在安全运营中实现降本增效的合理预期，以及用户内部创新研究的绩效牵引。
- ✓ LLM 提升了交互性并极大的增加了可解释性和推理能力，“安全大模型”的出现有助于安全价值可视化与用户体验两方面实现质的飞跃。
- ✓ “安全大模型”现阶段主要解决的是“人”的问题，聚焦在学习人的经验和模仿人的思维两方面。在单纯的攻防技术层面，尚未发现颠覆性的创新与应用。
- ✓ 高投入服务于高价值，现阶段由于我国数字化程度的不充分以及安全工作价值的不直观，高投入的 LLM 并不适用于所有安全场景。较高的投入产出比和较多的适用方向为“大模型+小模型”，即“安全调度官”。
- ✓ “安全大模型”的应用主要集中在攻防对抗智能化、威胁狩猎深度化、安全知识科普化和安全运营效率化 4 个方面。少量安全专项能力应用，集中在数据分析和代码应用 2 个方面。
- ✓ 具备较大模型规模（大于等于 60B）的“安全大模型”已经在数据分类分级中展现出强大的能力，有效降低了数据打标过程中的资源投入，基本替代了人力劳动，极大缩短了项目实施时间。
- ✓ 从全球范围来看，由于数字化建设程度和治理模式的区别，我国“安全大模型”的头部用户依然倾向于私有化部署方式，这就导致了供应商需要在“安全大模型”的资源投入、算力需求和应用性能之间作出权衡。但这也从侧面激励了供应商对预训练数据和指令调整的质量进行不断优化，来减少训练计算量和推理算力。
- ✓ 现阶段我国“安全大模型”的发展主要受制于算力资源紧张。长久来看则取决于安全法治、AI 基础研究和科技治理（开源、开放）的发展。

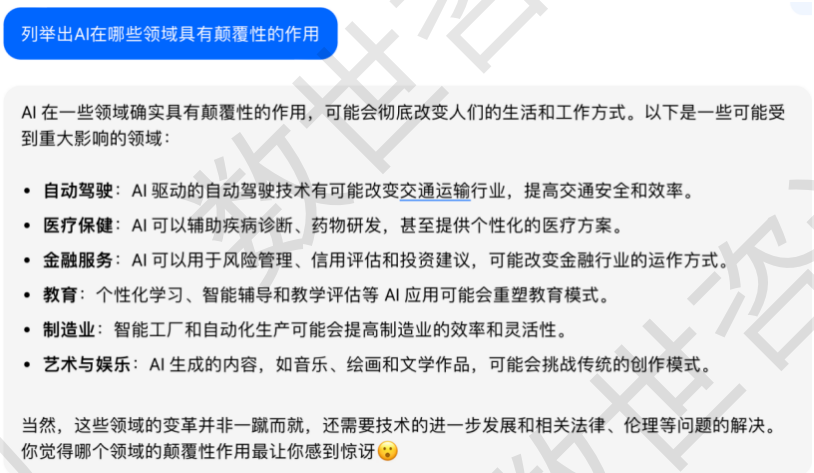
✓ “安全大模型”未来的发展方向是能解决多场景的关联任务，即“安全指挥官”，高级形态需具备多模态和多类型安全逻辑的“安全 AGI”能力。

第 1 部分 LLM 与数字安全的关系

如果说军事实力、经济实力是全球范围内大国博弈所使用的一对拳头，那么在世界格局动荡、百年未有之大变局中，抢占 AI（Artificial Intelligence，人工智能）高地就是霸主之争的战略目标，因为 AI 可以让拳头打得更准、打得更疼。

1950 年，艾伦・图灵提出了图灵测试的概念，指出如果一个机器能够在人类无法区分的程度上模仿人类的对话，那么就可以说这个机器具有智能。由此开始，专家系统、神经网络直到深度学习的兴起，一次次的推动着 AI 浪潮。

2016 年，AlphaGo 击败了围棋世界冠军李世石，AI 一石激起千层浪引起了全球范围内对 AI 应用的广泛讨论。发展至今，AI 已经利用机器学习、自然语言处理、计算机视觉、人机交互等核心能力对人类生产和生活方式进行了诸多颠覆性的改造，包括但不限于：（以下回答由“豆包”提供）

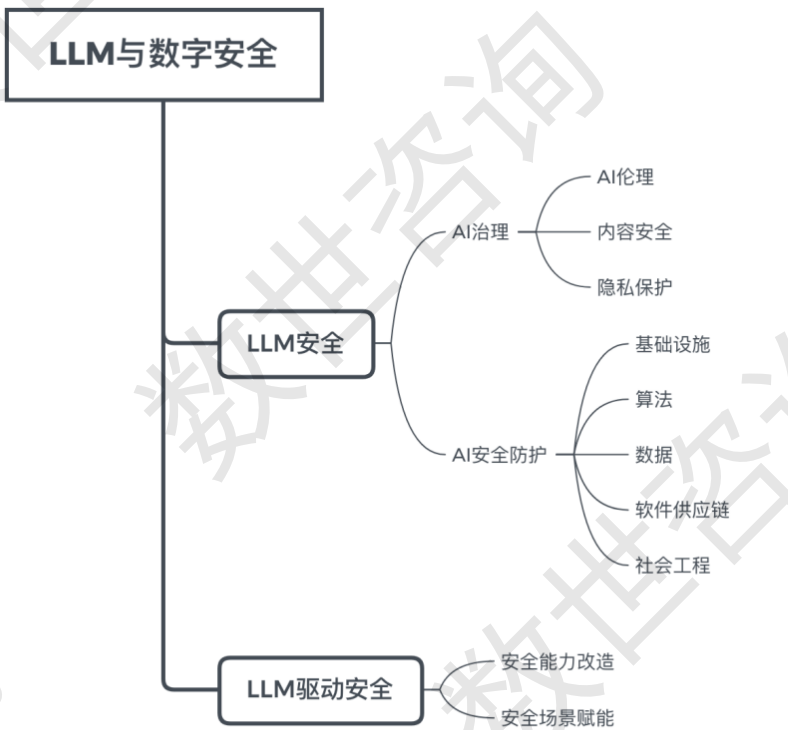


自 2022 年开始，在自然语言处理领域中的 LLM（Large Language Model）的突破性进展又一次给人类带来了巨大的冲击。以 ChatGpt3、DALL・E3、Sora 为代表的研究成果使得人类距离实现 AGI（Artificial General Intelligence，通用人工智能）更进一步。而最近 Devin 和 Figure-1 的发布，将“硅基生命”的未来再一次清晰

的展现在“碳基生命”眼前。

而 LLM 在赋能各产业的同时也在逐渐形成自身的产业生态，我国已经在国家层面确立了 LLM 发展的战略高度。同时，国内文心一言、通义千问、星火、盘古、云雀、混元、360 奇元、智谱清言、百川、Moonshot AI 等 LLM 也在不断的成长着，魔搭社区中更是展现出个人开发者在 LLM 应用上的百花齐放。

回到数字安全领域，有关 LLM 的研究和讨论主要包含 LLM 安全（AI 治理，AI 安全防护方向）和 LLM 驱动安全两类。



LLM 安全方面，除 AI 基础设施、软件供应链、社会工程与现有数字安全技术、

解决方案具有较高的匹配性外，其他方向的数字安全技术、解决方案均需根据 AI 工程和 LLM 原理进行升级以从根本上解决安全问题，才能体现出令人满意的效果。

数世咨询认为，在 AI 治理以及相关算法和数据安全这几个细分赛道，未来会诞生一批以 LLM 研发为核心能力的新兴数字安全供应商，并且成长为传统数字安全供应商的主要竞争者。

LLM 驱动安全方面，是数字安全供应商创新能力和技术沉淀的最佳战场，也是本次报告的核心关注点。利用 LLM 对安全能力进行改造以及各类安全场景赋能，是现阶段最具技术可行性和应用落地性的方式。

数世咨询认为，通过 LLM 的赋能，可以有效提升各类安全场景中的工作效率，适应并匹配新质生产力的特性要求，真正实现高质量发展和高水平安全的目标。更为重要的是，只有 LLM 驱动的数字安全才能对抗利用 LLM 的数字威胁。

1.1 LLM 的新兴能力

作为自然语言处理领域的重大突破，LLM 革新了人机交互、知识获取、内容生成方式，这些变革的产生，主要源于 LLM 新兴能力的涌现。而数字安全也正是利用这些新兴能力，改变了用户的体验。

LLM 的新兴能力是在小型模型中不存在但在大型模型中出现的能力，这是区分 LLM 与以前的 PLM（预训练语言模型）最显著的特点之一。主要表现为当模型规模达到一定水平时，性能显著高于随机情况。很多人习惯用复杂系统中出现的“涌现”现象来说明，也可以用物理学中“相变”（例如物体性状的转变）的概念来理解。

- 上下文学习（In-context Learning, ICL）：在提示中为语言模型提供自然语言指令和多个任务示例，无需显式的训练或梯度更新，仅输入文本的单词序列就能为测试样本生成预期的输出。
- 指令遵循（Instruction Following）：通过使用自然语言描述的多任务示例数据

集进行微调（指令微调或监督微调），大语言模型可以在没有使用显式示例的情况下按照任务指令完成新任务，有效提升了模型的泛化能力。

- 逐步推理（Step-by-step Reasoning）：对于小型语言模型而言，通常很难解决涉及多个推理步骤的复杂任务（如数学应用题），而大语言模型则可以利用思维链（Chain-of-Thought, CoT）提示策略来加强推理性能。具体来说，大语言模型可以在提示中引入任务相关的中间推理步骤来加强复杂任务的求解，从而获得更为可靠的答案。

1.2 什么是数字安全大模型

1.2.1 数字安全大模型定义

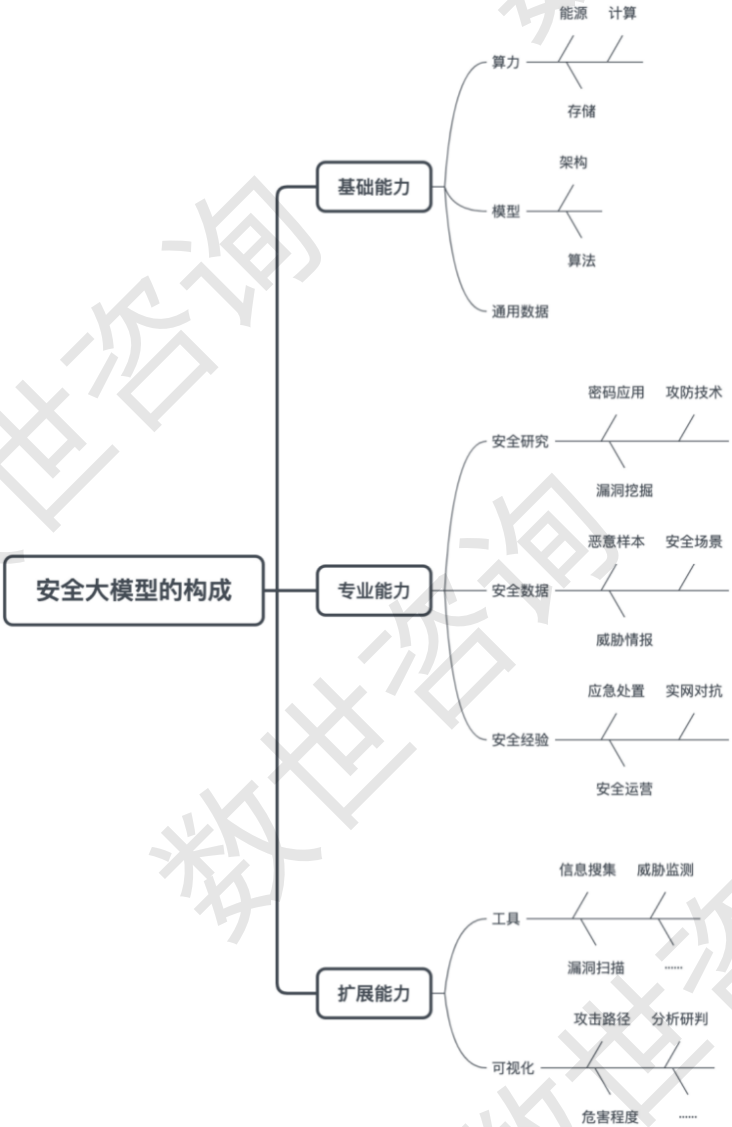
数世咨询认为，“安全大模型”（LLM 驱动的安全能力）是指 LLM 在数字安全领域的应用。

对于数字安全领域

- 专业能力包括：密码学、社会工程学和攻防对抗知识，以及专业级安全数据（威胁情报、安全运营信息等）和实网作战经验（实网攻防能力、应急处置能力等），专业能力的强弱决定了“安全大模型”的应用效果。
- 法律适应性（亦或者人类价值观对齐）特点：由于“安全大模型”和通用大模型在受众（如安全运营人员）、使用范围（如内网）方面的差异，在对齐调整上可以采取一些简单粗旷的方法（例如输入控制、敏感词过滤等）来减少整体资源投入，在现阶段甚至可以屏蔽非数字安全相关内容的输出。

1.2.2 数字安全大模型的构成

数字安全大模型由基础能力（通用模型，提供基础知识和 LLM 新兴能力）、专业能力（安全领域的知识和解决问题的逻辑）、扩展能力（利用辅助工具和可视化）三方面构成。



对于通用模型，现阶段大部分数字安全供应商都选择开源 LLM（LLaMa 和 Qwen 居多），少部分规模较大的企业会完全自主训练。对于一定规模之内（例如 10B）的 LLM 来说，自主训练的模型可能在安全知识的理解和问答上具备一定优势，因为预训练数据中安全数据的占比会比一般通用模型高。但模型规模较大（例如

60B 以上）时，因为用于预训练的安全知识以及安全数据自身数量有限，自主训练的模型优势几乎不复存在，反而会出现资源消耗过大的负面影响。除非数字安全供应商能够自主开发出针对安全知识和逻辑的专用算法、模型架构，否则只能使用开源通用 LLM。

对于安全数据，精准的样本、广泛的情报、具体的方法，这些都是在安全大模型预训练过程中决定“安全大模型”智商的关键所在，也是实现“安全大模型”在具体使用过程中提质增效的重点。

对于安全经验，掌握相关算法和架构的知识，才能训练或调整基础模型，基础模型解决了人机交互和逻辑推理的能力问题。具备数字安全实战经验的专家将自身经验转化成调整代码、学习实例，才能解决安全专业领域的问题。人的创造力和应用力解决安全大模型可持续发展和落地的问题。

对于工具使用，由于 LLM 知识的更新存在滞后性、分析数据缺乏实时性，要想在用户真实环境中实现较好的应用效果，协同工具的数量和融合数据的能力就是“安全大模型”效用延伸和价值提升的核心。

对于可视化，通过推理能力和上下文关联能力，不仅可以将威胁影响和风险程度分析并呈现，还可以增加安全分析的可解释性，提升处置建议的接纳程度。这种可视化能力将扭转安全工作价值难以衡量和可见性差的局面。

第 2 部分 卓越供应商&雷达图

由于安全大模型属于新兴技术，尤其在我国数字安全产业中暂未形成规模化市场，故本报告中所有市场方面的描述，是在综合考虑了已签单和试用项目对供应商的影响后，加权综合计算得出的。

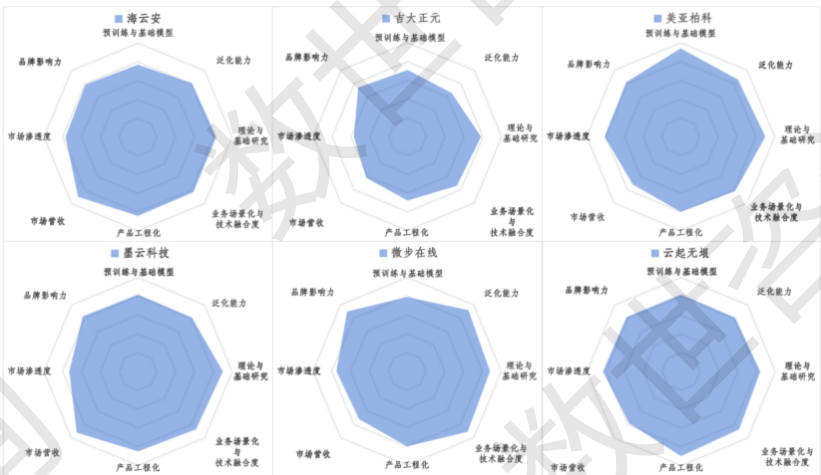
数世咨询认为，对于安全大模型，现在是一个探索大于应用的阶段，不适合以点阵图的方式为各供应商排序。本报告的核心目的是向行业用户展示安全大模型的可行性以及供应商在各方面的能力，故通过雷达图的形式将供应商所涉及的安全大模型的 8 个方向做出展示，以供行业用户参考。

- 预训练与基础模型：安全大模型对数字安全知识的储备和理解能力。
- 泛化能力：安全大模型处理数字安全问题的准确性。
- 理论与基础研究：供应商在 AI，尤其是 LLM 的资源投入和研发能力。
- 业务场景化与技术融合度：安全大模型对安全场景赋能的多样性。
- 产品工程化：安全大模型产品的成熟度和客户体验。
- 市场营收：供应商的市场执行能力。
- 市场渗透度：供应商客户的多样性和规模。
- 品牌影响力：安全大模型在行业和用户侧的认知程度。

本章所有关于供应商的展现顺序全部根据品牌中文首字母排序，入选本次报告的卓越供应商有：

2.1 专项类供应商

专项类供应商主要是指专注于某一方向，并且具备通过安全大模型提升产品能力的供应商。



- 海云安：代码检测方向应用。提升漏洞、安全隐患、隐私合规的检测准确度与效率。
- 吉大正元：数据安全方向应用。目前处于研究和试点应用阶段，重点面向关

键信息基础设施行业。

- 美亚柏科：业务数据分析方向应用。助力公安行业业务应用，提升查询、报告生成、案件辅助等效率。
- 墨云科技：智能攻防方向应用。赋能漏洞评估、修复建议、攻击模拟等。
- 微步在线：威胁情报的融合分析与扩展方向应用。赋能安全风险预测、研判、溯源等能力。除常规化应用外，还是首家通过网信办备案的公开安全大模型（https://x.threatbook.com），有效促动安全知识学习主观能动性，有助于提升个人安全研究能力。
- 云起无垠：模糊测试方向应用。解决模糊测试使用门槛高的问题，助力模糊测试易用化，极大地提高行业用户接受和使用度。除常规化应用外，还是首家中国开源安全大模型（https://github.com/Clouditera/SecGPT），有效促动数字安全开源领域发展。

2.2 综合类供应商

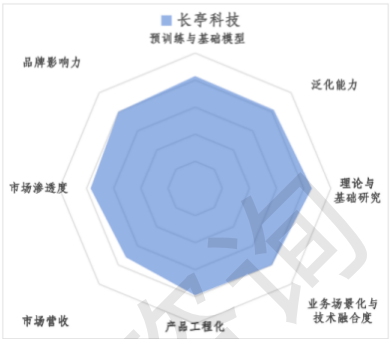
专项类和综合类在 8 个方向上的具体指标和评分权重不同，两种类别之间不具备比较性。

综合类供应商主要是指自身具备多产线研发能力，并且具备通过安全大模型赋能多个产品、或多个安全运营场景的供应商。

分析师标签

重保专业户，安全场景见多识广

数据分类分级初见成效



分析师标签

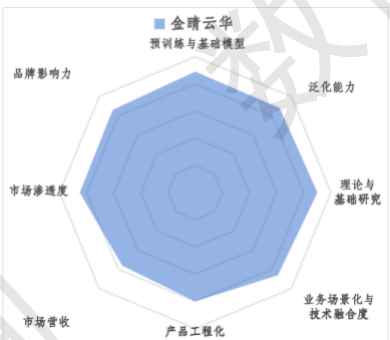
定位明晰，应用为先

工具多样，联动赋能

分析师标签：

基于“百业灵犀”LLM基础能力

依托新华三数字化能力，具备智慧城市的安全场景优势



分析师标签

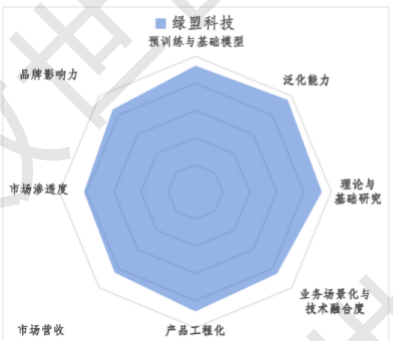
流量检测，火眼金睛

AI能力可与大厂掰手腕

分析师标签

工控安全赛道的尝鲜者

持续研究、蓄势待发



分析师标签

巨人背后的巨人

智能化攻击平台

分析师标签

国家级实网对抗能力

“八边形”战士

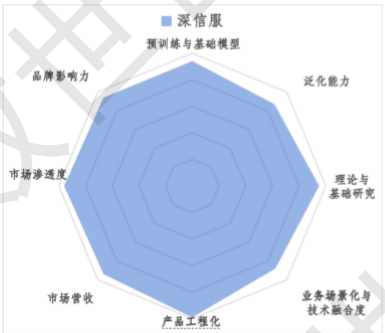
没有短板的木桶

分析师标签

APT猎手

双满分选手

AI同事“红衣”战斗力满格

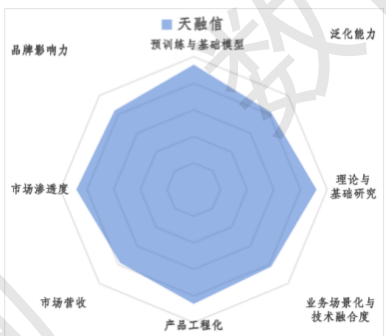


分析师标签

首家“安全大模型”商业化应用

领先的可视化与使用体验

客户“遍天下”



分析师标签

懂专业的AI客服，产品问题全知晓

“小天”赋能全线产品值得期待

分析师标签

专注安全运营管理

持续跟进多模态

2.3 安全运营解决方案

2.3.1 360 数字安全

- 场景问题

随着国际形势的不断变化，企业机构遭受网络攻击的强度和频率都在增加，用户面临的风险加大，需要有效的安全防护。但现实情况存在如下困境：首先，安全产品买了不少，但由于安全运营基本依赖人力，因此安全运营效率低下；其次，高级威胁日益增多，现有设备和人员安全能力不足，在高级威胁检测、溯源分析方面存在欠缺；第三，安全人员匮乏，高端安全专家稀缺，亟需有自动化的辅助工具提升人员能力。

- 总体方案

立足“小切口，大纵深”方法，从安全运营的智能化自动化入手，解决企业安全能力和运营效率低下的问题。构建以安全大模型为核心的智能体框架，结合现有安全工具库和安全知识库，实现强大的安全专家能力，打造智能化自动化运营体系。

➢ 安全知识问答

360 安全大模型系统支持对信息安全、计算机技术、开发语言、安全产品手册、政策法规标准、安全资讯、威胁情报和专业安全知识进行提问，可实现自动化交流。

➢ 辅助安全运营工作

360 安全大模型系统集成到安全运营工作平台的工作流节点中，为安全运营人员提供告警研判、日志解读、样本分析和响应处置等功能的自动化支撑。

➢ 安全威胁情报分析

360 安全大模型系统支持对 IP、域名、网址、样本、证书、病毒、漏洞和恶意组织等不同类型情报的智能问答分析。

➢ 威胁溯源分析

360 安全大模型系统通过统一调度大模型、情报和安全产品的能力，实现对告警的溯源分析，尤其在 APT 组织等高阶威胁溯源方面具有独特性。

- 用户价值

➢ 打造高端安全能力

360 安全大模型系统与已有或新建安全体系相融合，增强传统安全能力，全面实现安全能力的智能化升级。具体包括：代码理解能力、漏洞挖掘能力、日志解读能力、情报融合能力、威胁研判能力等。

➢ 创新安全产品体验

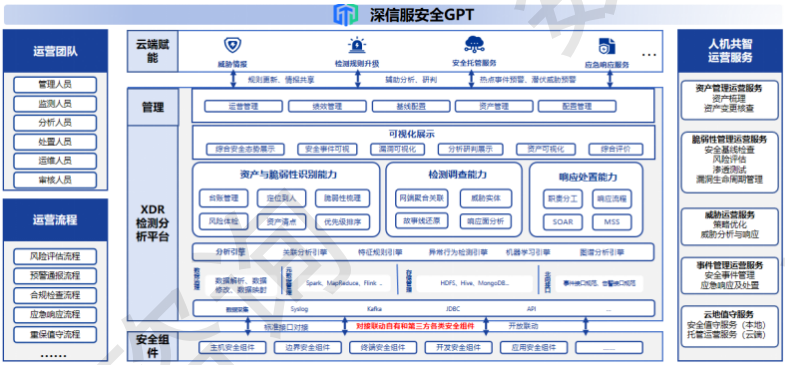
360 安全大模型系统将全面重塑安全产品，改变传统的操作模式，给用户带来全新的操作体验和安全效能。使用“知识助手”功能可以完成安全问答，使用“运营助手”功能可以完成辅助研判，使用“虚拟分析师”功能可以自动化分析研判，提出处置建议。

➢ 重塑智能安全运营体系

360 安全大模型系统颠覆了传统的安全运营理念，打通不同安全产品的界限，形成一站式看见、处置、溯源、报告等智能化安全运营能力。

2.3.2 深信服

深信服以“XDR+GPT”双擎，结合配套的安全运营服务，帮助用户构建安全效果领先的“3+1+1”安全运营体系：基于 XDR 重塑 3 项核心能力、结合 1 个安全 GPT 深度赋能、配套 1 类安全运营服务，实现“让实战对抗更高效，让运营工作更省心”的目标。



- 方案优势

➢ 极致降噪:实现海量告警 95%以上的降噪效果，形成少量精准的安全事件.

➢ 业界率先创新威胁定性能力:XDR 可实现对告警进行一键定性（业务误报、普通病毒、自动化攻击、定向攻击），帮助用户快速聚焦高风险的攻击威胁.

➢ 攻击故事线还原，构建高质量可视化故事线，精准还原安全事件的过程
可实现深度上下文的聚合分析，转换为用户能够理解的高价值安全事件，基于情境检测分析，精准还原黑客攻击链，大幅降低溯源难度，实现由局部检测转向整体分析的能力跃迁。

➢ AI 智能检测:聚焦高级威胁识别，检出率达到 95.7%，误报率仅 4.3%。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。

➢ AI 辅助运营: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%。

➢ AI 智能运营:安全的智能驾驶时代，安全 GPT 支持 7x24 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。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 的智能驾驶则可控制在 30s。

- 核心技术

➢ XStream 技术:深信服 XDR 创新采用 XStream 技术，通过整合多种 AI 技术，

包含自动接入引擎、威胁类型自动理解引擎、智能校验引擎，可大幅提升多源数据接入的效率。

- E+N+X 深度关联分析技术:基于 E+N+X 深度关联分析技术，按强关联、逻辑关联、弱关联等方式实现多源数据深度关联分析，有效发挥安全数据真正的价值。
- OPEN XDR 技术:基于 OPEN XDR 技术实现三方组件能达到与自有原生组件一致的安全效果，是 XDR 开放性、可生长性的关键技术。
- AI 大模型技术：利用自然语言与安全大模型进行交互，承载 80%安全运营操作，赋能初级安全工程师在 5 分钟内对单一高级威胁进行闭环；还能提供 7x24 小时实时在线自动研判、托管式值守，代替安全运营人员进行资产、漏洞的排查和管理等工作，提供秒级的闭环处置效率。
- ASM 技术：实现基于攻击视角的资产梳理、漏洞管理，可从攻击者可利用性、可触达性等多维度综合给出资产脆弱性优先级排序。

第 3 部分 有关市场现状的调研

由于安全大模型在我国属于初期市场，用户数量和应用场景、案例还不足以描绘出完整的市场概况，故本次调研，在市场层面主要涉及的问题有用户需求、用户行业分布、用户地区分布以及 2024 年预估市场营收等信息，通过这些数据分析，我们可以粗略的看到一些趋势。

各家供应商在 2023 年下半年开始集中在用户侧进行试用，少数供应商凭借强大的市场执行力与产品工程化直接获得了签约订单，有了真实场景的加持，给后期产品升级和迭代打下了坚实的基础，在与其他供应商竞争中占据先发优势。

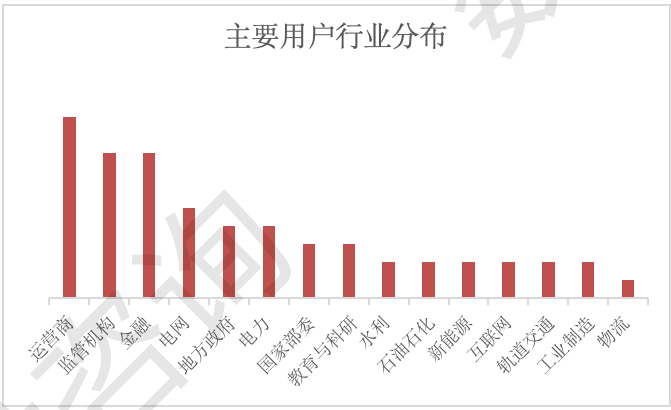
虽然安全大模型的战役在 2023 年已经打响了第一枪，然而 2024 年，才是各家供应商争夺市场占有率的真正起点。因为“两会”后明确的政策，安全大模型已经在政府部门和头部企业拥有了充足的预算，各家供应商也在 2024 年纷纷推出了正式对外销售的产品、服务。

2024 年，对于安全大模型以及通过安全大模型赋能升级的产品、服务，各家供应商预估总额可以达到 5.85 亿元人民币的市场规模。在这其中，预估最高销售额为 1.2 亿，预估最低销售额为 0.1 亿。需要注意的是，这里并不包括有关国家安全的部分，例如军事与军工研究等，并且这一部分完全具备充分的想象空间。

对于在安全大模型早期阶段显示出明确需求的用户，通常都具备一些共性，例如可以从政策层面获得一定的经营利益或者具有引领和创新研发的任务，当然也有一部分是因为自身业务以及数字化转型对安全大模型的迫切需求，还有一小部分是真正想通过安全大模型对自身安全能力赋能。

3.1 行业分布

通过数世咨询对行业用户和供应商的调研，统计出了主要用户行业分布：

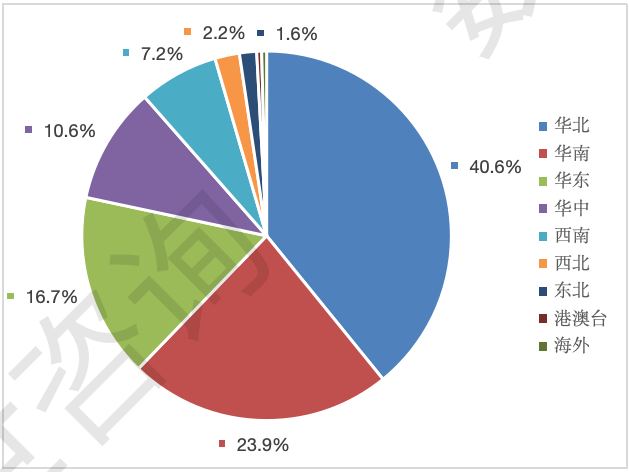


可以看到，运营商、监管机构、金融行业是现阶段安全大模型接受度最高的 3 个行业。这 3 个行业中，运营商和金融行业是安全运营工作开展较为成熟的行业，并且安全能力已经和业务能力紧密结合，安全运营已经成为保障核心业务的重要一环，安全大模型对安全运营工作的赋能是具备企业经营正反馈的。而监管机构则对安全数据的监测和分析具备极高的需求，通过安全大模型的赋能，监管机构可以使用更便捷的交互方式进行工作，可以纳管和分析更大范围的数据，可以更加便捷的生成相关报告等，直观的提高了工作效率。

还有 1 点，教育与科研行业一直是数字安全的弱需求行业，然而在安全大模型上却成为了主力选手。从这里也可以看到，LLM 在各行业的落地应用不仅仅是一句口号，已经成为了一个不争的事实。安全大模型在教育与科研行业的落地，也会进一步加强安全大模型的基础研究，真正形成产、学、研、用的正向发展链条。

3.2 区域分布

除了行业视角，我们还以区域视角进行了统计：



在安全大模型的主力需求用户中，华北地区以 40.6%的绝对优势占据首位，这与数字安全产业整体结构相似。华北地区主要由监管机构、运营商、科技企业总部以及教育与科研机构组成，安全大模型呈现出集中性需求。

华南与华东位列第 2 和第 3，华南主要集中在广东，华东依然沿长三角地区分布。华南作为技术创新和商业创新的高地，对于安全大模型的应用具备天然的吸引力。华东由金融行业为主，依托互联网以及智能制造业形成了安全大模型的集中需求。根据数世咨询的研究，华南和华东两地也是未来数字安全市场规模增幅潜力最大的地方。

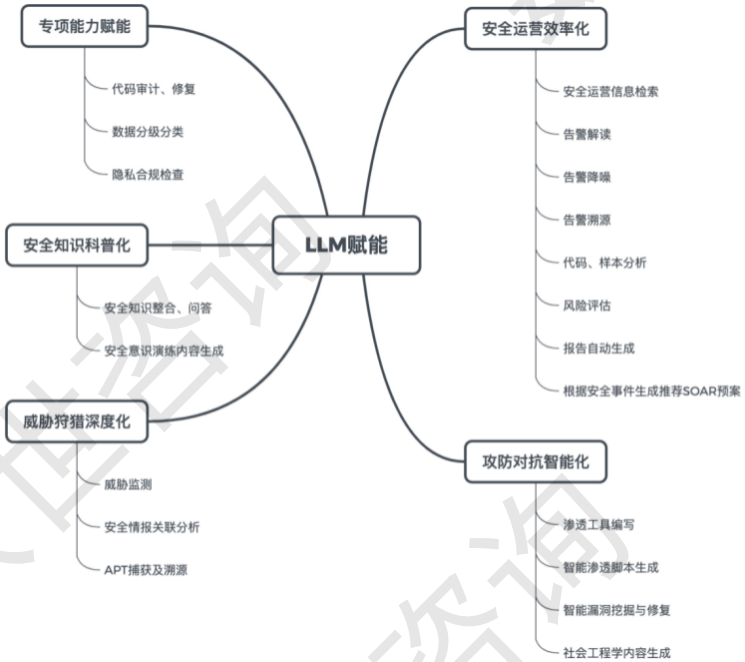
第 4 部分 有关能力应用的调研

现阶段，LLM 并不适用于全部安全场景或安全产品的赋能。一方面是因为训练安全大模型需要极大的投入，某些场景的安全能力提升结果与资源投入不成正比，并且有些安全场景的数据还不足以训练 LLM 以支撑其完成任务。还有一方面是因为 LLM 在不断的进化并且我国对 LLM 的基础研究还不够，LLM 的能力进化不是线性增长的，现阶段的困难对于下一版本的 LLM 可能不值一提，虽然我国已经有诸多本土 LLM，但大部分还是根据国外开源 LLM 修改而来，并且能力与 ChatGPT4.0 依然存在一定差距，而 LLM 基础研究的欠缺也直接导致了其在数字安全应用方面的限制。

4.1 能力应用

数世咨询通过调研发现，现阶段的安全大模型，主要集中在 5 个方向对安全能力进行赋能。

这 5 个方向主要包括安全运营效率化、攻防对抗智能化、威胁狩猎深度化、安全知识科普化和专项能力赋能，这 5 个方向也得到了用户侧的广泛认同，可以在实际生产环境中进行测试。



值得注意的是，7B 以上规模的安全大模型就具备解决安全问题的能力，而数据分级分类与隐私合规检查普遍需要 60B 以上的模型规模才具备理解和推理业务数据的能力。这种差异主要由处理对象的不同而产生，安全数据分析主要为代码，而数据安全业务主要需求对人类语言的理解，而 60B 以上规模的 LLM 才可以达到基础要求。

现阶段的安全大模型规模主要集中在 7 至 15B、30 至 40B、60 至 70B 三个区间，根据不同的安全场景匹配不同规模的安全大模型，真正做到性能与效果的平衡。

与此同时，数世咨询还发现了现阶段对于安全大模型应用的共性难点，大致可分为两方面，一是应用的难点，一是落地的难点。

4.2 应用的难点

- 应用成熟度：国内大模型技术虽然发展迅速，但与国际先进水平如 OpenAI

的 ChatGPT4.0 相比，仍存在差距，需要持续研究和突破。安全领域的应用技术尚未完全成熟，存在技术和成本挑战。

- 数据可用性：尽管拥有大量安全数据，但直接适用于大模型训练的数据比例较小，需要大量人力进行数据优化、清洗。
- 算力资源：训练大模型需要巨大的算力资源，而当前 GPU 资源受限，导致训练成本显著提高。
- 价值观偏见：大模型可能放大或曲解训练样本中的偏见，需要确保生成内容合法合规且符合主流价值观。
- 数据安全与隐私保护：需要在保证训练效果的同时确保数据安全性和隐私保护。
- 应用技术成熟度：安全领域的应用技术尚未成熟，存在技术和成本挑战。
- 模型安全：需要防御恶意数据对模型可靠性和稳定性的威胁，保护模型不被攻击者通过询问获取内部结构和功能，提高模型鲁棒性防止对抗样本导致的错误输出，有效检测和防范后门攻击确保模型输出的正确性。

4.3 落地的难点

- 用户认知：部分用户对大模型技术存在误解，有的神化其能力，有的则低估其潜力。
- 产品结合：安全大模型需要与现有安全产品深度结合，但用户对于购买新产品的驱动力不足。
- AI 工程化能力：用户需要具备一定的 AI 工程化能力，尤其是在增量训练或微调方面。
- 私有数据和网络环境限制：企业私有数据和网络环境限制导致无法在外网训练私有数据，且难以接受 SaaS 化部署方案。
- 算力成本：算力资源昂贵，增加了企业的成本负担。
- 法规与合规性：遵守各国、各行业对人工智能的法规与合规要求，保护用户隐私和数据安全。
- 第三方权威机构：缺少大模型领域的第三方测评机构和标准，缺乏具有公信力和号召力的联盟组织统筹行业标准、规范，亟需第三方对接广泛应用和交

流活动。

第 5 部分 创新与实践案例

5.1 奇安信安全大模型赋能北京银行安全运营

5.1.1 应用环境与需求

近年来，北京银行积极拥抱数字化转型，致力于构建数字银行 4.0 时代，在数字化环境建设方面取得了显著进展。在 2021 年年报业绩说明会上，北京银行公开表示，通过三年时间，推动北京银行数字化转型达到同业领先水平，建成国内领先的数字银行。

数字化环境方面，北京银行采取了多项举措。首先，提出了“数字京行”战略体系，并坚持以数字化转型统领发展模式、业务结构、客户结构、营运能力、管理方式等五大转型。然后，成立了数字化转型战略委员会、金融科技委员会以及北京市首家金融企业科学技术协会，以完善数字化转型的顶层设计。这些措施共同推动了北京银行在数字化环境建设方面的快速发展。

主营业务方面，北京银行涵盖了个人银行业务、公司银行业务、资本市场业务、国际业务和小微企业金融等五大板块。个人银行业务包括开户、存取款、信用卡、个人贷款、投资理财、电子银行等多种业务，为客户提供全方位的金融服务。公司银行业务则涵盖了资金管理、贸易融资、结算清算、投资理财、企业贷款等，以满足企业客户的多元化金融需求。资本市场业务包括股票交易、基金销售、债券发行、期货交易等，为客户提供全面的投资咨询和交易服务。国际业务则包括进出口贸易融资、信用证业务、外汇买卖、对外投资融资等，助力客户拓展国际市场。此外，北京银行还致力于服务小微企业，提供小额贷款、融资租赁、小额保理、小额担保等多种金融业务，支持小微企业的发展。

北京银行以“数字京行”战略体系为引领，锚定数字化转型方向，启动数字化转型“新三大战役”，通过全面数字化赋能，实现规模、结构、效益、质量的均衡稳健发展。具体为三步走策略：

- 建设并补全网络安全基础能力，实现网络安全能力内生，达到安全能力全覆盖、无死角；
- 采用集约化架构，建设由多源威胁检测、安全大数据归集、威胁计算与智能化分析研判、安全事件自动化响应处置组成的四级网络安全运营系统，建立安全事件自动化检测与响应处置闭环管理流程；
- 利用 GPT 人工智能技术，提高安全告警分析研判效率，加大安全事件自动化、智能化处置比例，积极探索将人工智能技术与网络安全防御体系全面融合，实践一条网络安全架构内生、网络安全能力互操作、自适应发展的道路。

5.1.2 总体设计

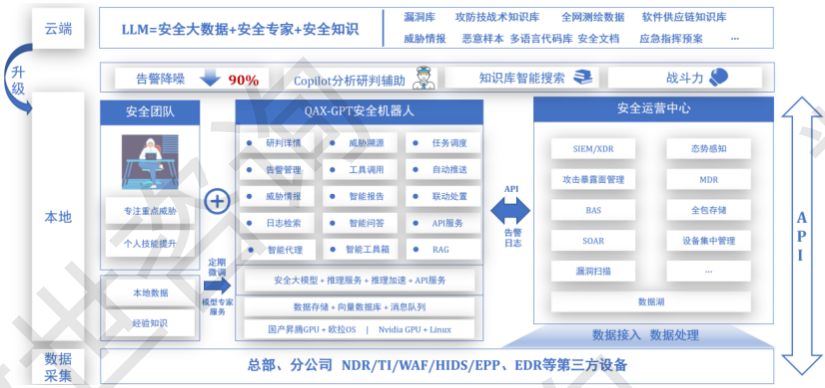
2023 年开始，北京银行不断加强以大模型等 AI 技术为核心的新一代金融智能基础能力，包括成立“金融智能化创新实验室”，聚焦金融大模型体系构建和创新应用；持续升级“京智大脑”建设，构建“智慧服务、智慧决策、数字员工”三大应用体系，累计上线智能服务 150 余项，智能决策模型 400 余个。

在网络安全方面，北京银行的安全监控原始告警已经达到每年大几千万条，即使通过态势感知平台筛选后，每天仍然有几万条告警信息需要安全团队手工处理。2024 年北京银行会进一步扩大信息源的接入，细化颗粒度，随之而来的信息量还会增长 3－5 倍，安全团队手工处理告警的压力很大。

此外，态势感知平台是依靠安全工程师的历史经验来撰写规则、处理信息的。对于每一类新的安全告警，都需要撰写新的规则，算上数据建模和测试验证一般都需要两三个月的时间才能上线。而北京银行数字化转型进程，处处要求安全同步规划，同步建设，当前的告警处理节奏显然会面临极大的压力。

这就是北京银行和奇安信合作共建 QAX-GPT 安全机器人的起源，当前阶段北京银行和奇安信想通过这个机器人来逐步取代初级安全运营工程师的工作任务，顶住源源不断涌过来的告警数据。机器人要能够理解告警内容，还能主动调度后台的各种自动化能力。再往后，还能够学会用好 AI 打败坏 AI，通过 QAX-GPT 安全机器人打击暗黑 AI 生成的虚假信息，还能够在北京银行全行实现完全个性化

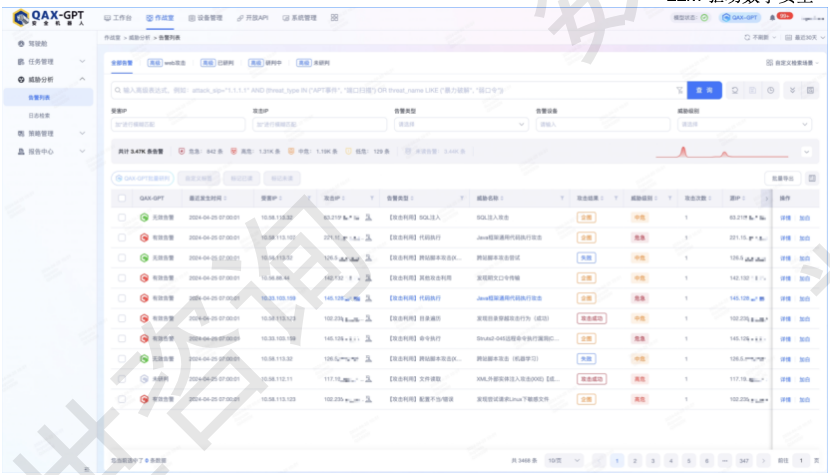
的安全意识文化教育和实战化训练，以及完全个性化的安全技术支持的智能服务。下图为项目建设总体架构：



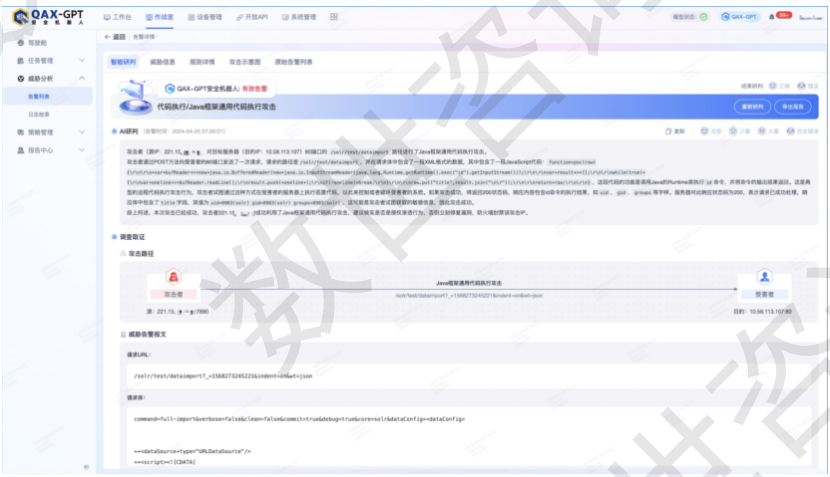
北京银行利用 GPT 等人工智能技术打造智能化安全运营新模式，构建一个能够自动化处理告警和提供智能建议的系统框架，提升网络安全防护能力。QAX-GPT 安全机器人一方面可以用大模型分析能力对海量告警进行全天候分析，另一方面可以将行内的知识沉淀为知识库，成为安全运营团队的资产。

5.1.3 落地与应用

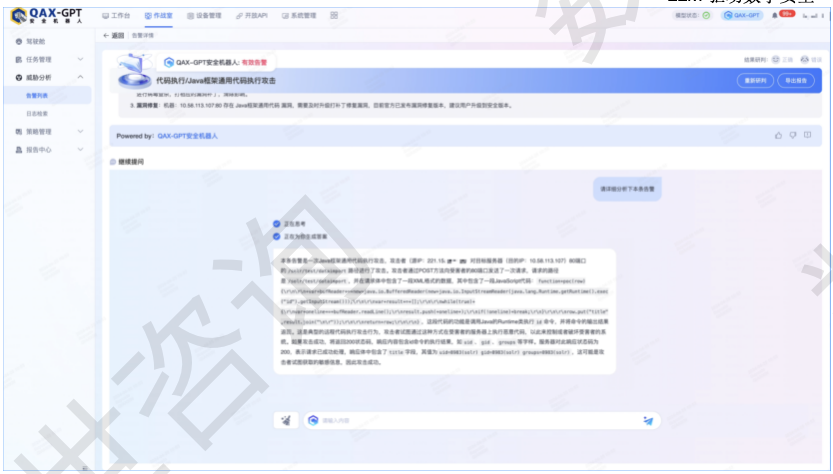
QAX-GPT 安全机器人能 7*24 小时全天候快速研判威胁告警，并从中甄选出真实有效的威胁告警，通俗易懂地全面呈现详细的解读过程。支持查看告警的详细智能研判信息，包括攻击的原理分析、过程分析、结果分析、处置建议等。如对攻击过程存疑，支持查看 QAX-GPT 安全机器人解读的详细攻击过程和取证信息。针对威胁可进一步对机器人提问，探索与该威胁相关的详细信息。



QAX-GPT 安全机器人平台智能告警分析列表界面



QAX-GPT 安全机器人平台智能告警分析详情界面

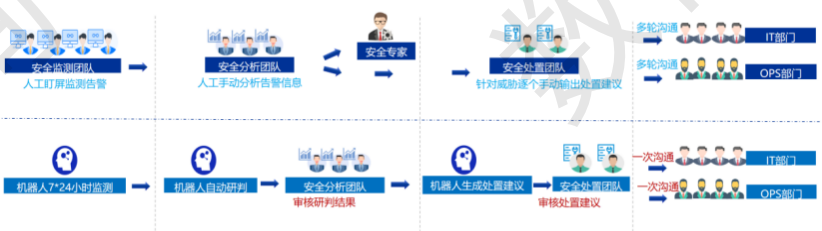


QAX-GPT 安全机器人平台智能问答界面

5.1.4 应用效果

北京银行为解决海量告警导致的告警监控疲劳问题，创新性地运用 GPT 人工智能技术处理网络安全设备产生的告警数据，人机配合大大超出了纯人工告警监控、分析研判的工作效率。同时，QAX-GPT 安全机器人对主要网络威胁的分析研判准确率，基本达到了初级和中级网络安全运营人员水平。

这些研究共建达成的成果，是双方全新合作模式的前瞻性探索，也是金融行业网络安全能力的数字化和智能化发展的宝贵经验。下一步，北京银行和奇安信将共同努力，在态势感知平台安全运营流程中加入 QAX-GPT 安全机器人，采用“人机混编”（下图）新模式，大幅度提升安全运营人效。



5.2 360 安全大模型赋能金融企业安全运营

5.2.1 应用环境与需求

XX 省 XX 银行是国有大型股份制商业银行，为全国的农村和小微企业提供全方位的金融服务。为了保护数据及网络安全，企业部署了很多传统网络安全产品用于保护基础的网络安全，在运营的过程中发现仍然面临一些痛点问题，首先，产品间缺少联动，难以形成防御体系；其次，高级威胁日益增多，现有人员和设备的安全能力不足；第三，安全运营基本以来人力，安全运营效率低下；另外还有行业普遍存在的安全人员，尤其是高级别安全专家稀缺的问题。

针对当前面临的产品协同困难、安全能力不足、运营效率低下、安全专家缺乏等痛点问题，客户提出积极拥抱人工智能技术，从“AI+安全运营”入手，解决网络安全运营面临的能力和效率问题。

5.2.2 总体设计

该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”四个方面核心战法，就企业当前关心的产品协同困难、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。

360 安全智能体由 360 安全大模型、记忆存储、任务编排引擎、任务生成引擎、监督评测引擎、指令调度引擎、执行反馈 7 个核心模块组成。安全智能体通过 API 接口灵活准确地调用各类安全工具库、安全知识库，完成多样化的安全运营任务。

- 安全大模型：360 安全大模型打造类脑框架，类比人类大脑划分为语言中枢、规划中枢、判别中枢、道德中枢和记忆中枢。五大中枢通过联动协作，实现专家级安全分析能力。
- 智能体配套模块：智能体以大模型为核心，辅以“记忆存储、任务编排引擎、任务生成引擎、监督评测引擎、指令调度引擎、执行反馈”等关键组件,构成一个功能全面的自治系统。
- 安全知识库、安全工具库：智能体与安全知识库、安全工具连接起来，就像组装一台复杂的数字机器人一样，让整个数字人运转起来，完成特定的复杂任务。

5.2.3 落地与应用

- 辅助工具

360 安全大模型系统支持对安全知识、产品操作、威胁情报、态势报告、安全法律法规等问题进行提问，可实现自动化交流。辅助安全运营专家查找知识、使用工具。

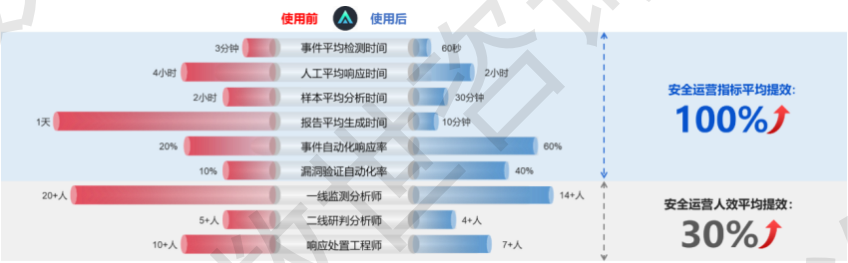
- 运营助手

360 安全大模型系统集成到态势感知平台的工作流节点中，为安全运营人员在告警研判、日志解读、样本分析和响应处置等功能提供帮助。

- 运营自动化

360 安全大模型系统通过调用安全知识、工具，充分发挥检索和工具的增强能力，实现对告警的自动化溯源分析，尤其在 APT 组织等高阶威胁溯源方面具有独特性。

5.2.4 应用效果



5.3 深信服安全大模型赋能政府行业安全运营

5.3.1 应用环境与需求

某国家部委的数字化环境呈现出高度复杂和庞大的特点，涵盖了约 300 个系统和 30 种安全设备，在日常运营中产生了海量数据。安全部门每天需要负责处理和管理一天内产生的数亿条日志和数万条安全告警。安全事件管理存在滞后发现、事件处理效率低下（预计耗时 7~8 小时以上）以及安全运维报告的编制周期过长（至少 1 周时间）等挑战，使得该客户对于安全运营的效率和效果有着迫切的提升需求。

客户需要一个统一的平台，能够整合现有的 300 多系统以及众多安全设备的数据，实现一站式登陆和集中管理。考虑到人工审查能力的有限性，客户期望将每天产生的安全告警数量大幅缩减，将平均每次安全事件的处理时间缩短至 0.5 小时左右，希望安全运维总结报告能够自动导出，且立即可取。

5.3.2 总体设计

深信服安全 GPT 建设的总体设计基于已有安全运营平台，实现生成式人工智能模型的应用升级，提升安全运营的智能化、自动化水平。

- 安全检测大模型：通过对网络流量的深度分析，及时发现潜在的安全威胁，增强代码混淆、编码绕过类的 0day/1day 攻击检出能力。
- 辅助运营大模型：基于自然语言交互的安全运营助手，快速了解漏洞数量、类型和严重程度等信息。在安全事件研判环节，通过按键触发安全辅助，快速闭环运营工作，包括告警解读、情报查询、事件分析等。
- 智能运营大模型：通过智能运营大模型的部署及运行，实现全网 24 小时人工智能自动化值守。该模型能够对所有告警进行逐一研判，并利用安全 GPT 进行自主告警分析，协同各类安全设备进行联动处置。
- 安全 GPT 方案涉及一台检测大模型服务器、两台运营大模型服务器。流量探

针实时采集流量，送入安全运营平台；经过该系统的基础过滤，将 HTTP 流量送入检测大模型；检测大模型检测分析后，将结果送回安全运营平台进行统一运营。将客户安全运营平台上的部分数据以 API 调用的方式按需传输至运营大模型，包括告警统计信息、事件统计信息、流量审计数据、资产属性数据（责任人名称、联系方式及漏洞）。

5.3.3 落地与应用


高级威胁检测


对话式辅助运营助手


全天候自主值守机器人

深信服安全 GPT 结合安全运营平台，构建智能调度控制中枢，实现与当前领先的安全系统（如 NDR、EDR、威胁情报、安全沙箱等）的整合，赋能网络安全，并能实现对其 API 接口的智能调度。支持基于问答式的网络安全分析模式，支持对日均≥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。

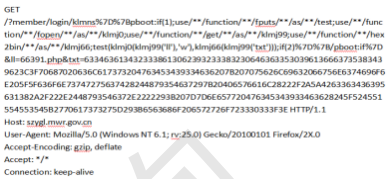
2023 年底，本地化检测大模型、运营大模型均已在客户侧实现实际业务环境上线部署，24 年初完成实际环境第三方安全产品接入和验证测试。

5.3.4 应用效果

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。

首次测试【基础】（2 月初）		二次测试【独报】（3 月初）	
测试持续时间	20 天	测试持续时间	一周
深信服安全 GPT 检出总数	2.3 万个	独检条数	8000 余条
正报率> 98%		独报告警占比	＞80%
二次测试【精准率】（3 月初）		**二次测试检出类型**	**占比**
深信服安全 GPT 告警总数	1.6 万条	信息泄露	＞50%
人工研判（抽检率>15%）		高混淆类网站攻击	＞20%
深信服安全 GPT 精准率	＞95%	高级代码注入类	＞15%
深信服安全 GPT 误报率	＜4%	其他	＜5%

在该国家部委测试过程中，由检测大模型在业务环境中发现一例高价值的webshell 上传攻击案例。该攻击 payload 采用了函数调用替换、十六进制码混淆绕过。通过构造一个名为 GC 的类，利用 PHP 的 eval 函数执行传入的参数作为代码。其中，参数$Hy5F7 经过混淆后为"echo error303"。在实例化该类时，传入的参数为$_REQUEST['pass']，即获取 HTTP 请求中名为'pass'的参数值。最终，该 payload 会输出"error303"。经过海量安全语料预训练的大模型，基于超强的安全自然语言理解的能力，能够通过安全逻辑理解精准识别该攻击。



运营大模型方面，对话式辅助运营，风险研判和处置闭环实现再次迭代提速，部里日常安全运营时效提升 20%以上。告警 GPT 自主研判，告警降噪 99.8%以上，高级安全运营人员精力充分释放，更聚焦于高级威胁研究及严重威胁闭环，全面告别事务性工作（资产统计、报告总结等），工耗降低 25%，部里安全防护质量显著提升。补充夜间安全监测研判处置力量，实现部里全天候 7*24 小时无间隙安全值守。

第 6 部分 趋势与挑战

新质生产力，无疑是现在最火热的词。习近平总书记指出，“发展新质生产力是推动高质量发展的内在要求和重要着力点”。新质生产力主要指战略性新兴产业和未来产业，AI 尤其是现在的 LLM，必定是当之无愧的新质生产力。

2024 年政府工作报告中明确指出开展“人工智能+”行动，强调了人工智能在各行各业的落地应用，通过基于大模型、大数据、大算力技术，提升产业自动化水平，实现降本增效。这显示出政府对于人工智能技术的深度理解和高度重视，也预示着未来人工智能将在我国经济社会发展中发挥更加重要的作用。

安全大模型同样属于 AI 在行业的落地应用，通过我们的调查与研究发现，安全大模型已经具备了一些明显的发展趋势：

- 能力提升：随着 LLM 算法和预训练数据的质量不断提升，安全大模型的推理及理解能力同样不断提高。未来的模型将会具备持续自我学习的能力，能够在不断积累新的知识和经验的过程中自我优化和更新，智能化能力可能超过高级专家。
- 能力重塑：越来越多的安全场景将会被赋予 LLM 的能力，几乎所有的安全产品都会通过 AI 重塑，安全运营的逻辑与实际操作过程将会迎来颠覆性的变革，不同规模组织间的安全保障能力进一步被拉开差距。
- 安全 AGI：现阶段的安全大模型基本都只能解决某一场景（例如流量分析、告警降噪等）的安全问题，并且都是基于文本（包括代码）的输入和输出。未来安全大模型将会实现安全 AGI，也就是具备解决多场景关联性问题的能力，并且具备多模态的输入与输出。例如通过资产探查绘制网络拓扑图，根据拓扑图生成安全控制方法与流程。例如通过摄像头监控机房人员的操作，发现风险事件及时提醒等。

政策的牵引、资源的偏移、人才的聚集、市场的需求，这对于安全大模型来说是“泼天的富贵”。但机遇总是与挑战同行，如何创造出符合市场需求与用户预期的安全大模型，至少有以下几点需要行业共同思考：

- 高可靠性：由于安全工作的特殊性，安全事件一旦发生，将可能对组织造成灾难性的打击。目前为止，对于安全大模型的输出还不能保证 100%准确率，甚至可能出现幻觉。试想安全大模型在一次攻击事件的研判中出现幻觉，将攻击视为正常访问，这样的风险事件是安全工作完全不能接受的。怎样保障安全大模型推理的准确性，以及怎么控制幻觉的产生，是安全工作中必须解决的问题。
- 可解释性：可解释性也就是模型透明性，应该有一种方式可以描述模型本身、预期影响和潜在偏差，从而使 AI 产生的流程和决策更好的被人类所理解和接受。例如解释安全大模型为何将特定的流量标记为潜在威胁，从而使人们对检测过程更加信任。例如解释安全大模型对不同 ID、不同网络的行为分析不会存在偏见，来保证公平性以及阻止意外事件发生。
- 评估标准：众所周知通用大模型是有一系列评估标准的，但是这些标准对于安全大模型来说并不完全适用，或者说并没有依据安全大模型所需要具备的能力进行针对性的评估。对于安全大模型，可以从两方面进行评估标准的研究与制定。一是预训练阶段获得对知识和概念的理解和表述，即“安全基础评估”。一是微调阶段将模型与用户的意图对齐，赋予模型按照指令行动的能力，即“安全任务评估”。