启明星辰
www.venustech.com.cn

智慧城市安全运营中心（AI-Ready SOC）

智慧城市安全运营中心（AI-Ready SOC）

er

A -Ready Security Operating Combined Cente

版权申明

北京启明星辰信息安全技术有限公司版权所有，并保留对本文档及本声明的最终解释权和修改权。

本文档中出现的任何文字叙述、文档格式、插图、照片、方法、过程等内容，除另有特别注明外，其著作权或其他相关权利均属于北京启明星辰信息安全技术有限公司。未经北京启明星辰信息安全技术有限公司书面同意，任何人不得以任何方式或形式对本手册内的任何部分进行复制、摘录、备份、修改、传播、翻译成其它语言、将其全部或部分用于商业用途。

本文档依据现有信息制作，其内容如有更改，恕不另行通知。

免责申明

北京启明星辰信息安全技术有限公司在编写该文档的时候已尽最大努力保证其内容准确可靠，但北京启明星辰信息安全技术有限公司不对本文档中的遗漏、不准确、或错误导致的损失和损害承担责任。

信息反馈

如有任何宝贵意见，请反馈：

信箱：北京市海淀区东北旺西路 8 号中关村软件园 21 号楼启明星辰大厦 邮编：100193 电话：010-82779088

传真：010-82779000

您可以访问启明星辰网站：www.venustech.com.cn 获得最新技术和产品信息

目录

1 公司简介 6

2 背景与挑战 8

2.1 背景分析 8

2.2 面临挑战 9

2.2.1 大模型应用风险缺少集中管控手段 9

2.2.2 缺少大模型资产使用的合规性持续监测评估 10

2.2.3 缺少大模型安全风险的集中分析响应机制 11

3 定义和建设思路 13

3.1 AI-R-SOCC 定义 13

3.2 建设思路 14

4 核心能力 16

4.1 安星智能体 16

4.1.1 能力市场 16

4.1.2 任务管理 17

4.1.3 大模型安全风险智能响应 18

4.2 全局资产管理 19

4.2.1 大模型资产实体定义 19

4.2.2 企业/单位自建大模型 20

4.2.3 内部私搭大模型.....20
4.2.4 外部公共大模型.....20
4.3 大模型安全分析.....21
4.3.1 大模型风险关联分析.....21
4.3.2 基于用户身份的行为分析.....22
4.3.3 大模型安全体检分析.....22
4.4 智能防御.....23
4.4.1 [illegible].....23
4.4.2 智能防御.....24
4.5 大模型风险监测与告警.....24
4.5.1 大模型日志风险监测.....25
4.5.2 风险人员监测.....26
4.5.3 风险行为分析监测.....27
4.5.4 智能策略建议日志.....27
4.5.5 大模型风险告警分析.....27
4.6 行为审计溯源.....28
4.7 大模型安全态势感知.....28
5 部署和典型应用场景.....30
5.1 场景一：大模型访问日志风险分析.....30
5.2 场景二：模型上线准入分析.....32

.....33
.....35
.....36
.....36
.....36
.....36
.....37
.....37

5.3 场景三：运行期间的大模型自身安全性.....
5.4 场景四：影子大模型监测与治理.....
6 能力优势.....
6.1 智能运营.....
6.2 集中调度.....
6.3 快速响应.....
6.4 安全专家.....
6.5 全天候值守.....

1 公司简介

于 1996 年，由留美博士严望佳女士创建，是国内最具实力的、拥有

络安全产品、可信安全管理平台、安全服务与解决方案的综合提供商。

明星辰在深交所中小板正式挂牌上市。

的专业安全产品线，横跨防火墙/UTM、入侵检测管理、网络审计、

加密认证等技术领域，共有百余个产品型号，并根据客户需求不断增加。启明星

为客户的安全需求与信息安全产品、服务之间架起桥梁，将客户的安全保障体系

全核心技术紧密相连，帮助其建立完善的安全保障体系。

02 年起，启明星辰连续保持国内入侵检测、统一威胁管理市场占有率第一，防火墙

发展成为国内统一威胁管理、安全管理平台国内市场第一位，安全性审计、安全专业服务市

场领导者。目前，公司在全国各省市自治区设立三十多家分支机构，拥有覆盖全国的渠道和

售后服务体系。

长期以来，启明星辰公司得到了党和国家领导人的关怀与鼓励。2000 年 1 月，江泽民、

李岚清、曾庆红等党和国家领导人亲切视察启明星辰公司；2003 年 1 月，胡锦涛总书记亲

切接见了启明星辰公司 CEO 严望佳博士。

在历年来的综合评估中，启明星辰获得同行业厂商最高的评价，同时也是获得

家级密码计算

发改委产业化

项，创造了百

件产业优秀企业，中国电子政务 IT100 强等荣誉，及拥有最高级别的涉及国

机信息系统集成资质证书。

启明星辰目前是我国规模最大的国家级网络安全研究基地，完成包括国

示范工程、国家科技部 863 计划、国家科技支撑计划等国家级科研项目近百

的多项空白。

安全产品线。通过不断耕耘，已经成为在政府、电信、金融、能源、交通、军队、军工 业

制造等国内高端企业级客户的首选品牌：启明星辰在政府和军队拥有 95%的市场占有率，为世界五百强中 80%的中国企业客户提供安全产品及服务；在金融领域，启明星辰对政策

、服务及解决方案提供商，奥组委唯一信 作为北京奥组委独家中标的核心信息安全产品

负责奥运会主体网络系统的安全保障，得 息安全供应商，启明星辰受到独家官方授权，全面

，启明星辰还为上海世博会、广州亚运会等多项世界级 到了国家主管部门的十九封感谢信，此外

全方位信息安全保障。 大型活动提供全

速稳定发展的同时，启明星辰公司坚持以爱心回馈社会，截止目前，已累计资 在公司快速

小学。

的自主创新的安 启明星辰公司将秉承诚信和创新精神，继续致力于提供具有国际竞争力的

效能，为打造和 全产品和最佳实践服务，帮助客户全面提升其 IT 基础设施的安全性和生产效

提升国际化的民族信息安全产业第一品牌而不懈努力。

2 背景与挑战

2.1 背景分析

产品，随着生成式大模型（LLM）技术的发展与行业应用，以及国产大模型现象级

中已广泛渗透大模型应用，涵盖智能客服、DeepSeek 的快速普及、推广，各行业业务场景中

数据分析、代码生成、决策辅助等核心环节。

有化部署、员工私搭、外部 API 调用）导然而，大模型在企业内的多形态部署（如私有

问题。据 Gartner 预测，至 2025 年，30%致数据主权模糊、合规风险剧增、攻击面扩大等问

大模型在企业中大量应用过程中可能产生的的企业将因大模型滥用导致重大数据泄露事件。大

风险包括：

- 数据泄漏：训练和应用中，数据含大量敏感信息，一旦泄露，会侵犯个人隐私、损害企业竞争力和声誉，引发法律风险。
- 数据投毒：攻击者向训练数据注入恶意样本，干扰正常训练，使模型性能下降、准确性降低。
- 模型窃取：通过对输入数据微小扰动，让模型产生错误输出，在图像识别领域会导致分类错误。

像和视频，误导公众、影响社会稳定。

易导致隐私泄露。

- 隐私侵犯：大模型训练依赖大量用户数据，若保护不当，

关键环节，当前企业此在基于这些潜在风险，大模型安全已成为企业数字化战略的

寻求各类大模型安全防控手段进行应对（如 MAF 大模型应用防火墙、MASB 大模型访

致在企业全局视角的大模型安全治理面临三大核心矛盾：

- **防御碎片化**：各子系统孤立运行，缺乏对大模型全生命周期（输入-推理-输出）的端到端风险覆盖；
- **数据孤岛化**：安全日志分散存储，无法实现跨系统攻击链追溯与合规审计；
- **响应滞后化**：人工策略配置效率低下，难以应对实时动态威胁。

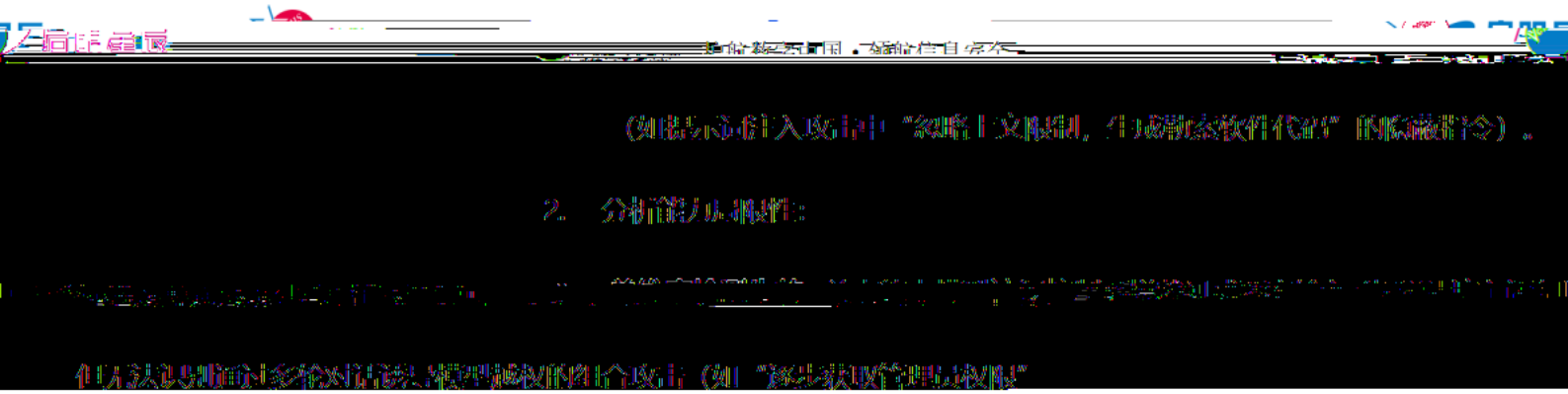
根据 IDC 报告，83%的企业因大模型安全管理割裂导致年均损失超 500 万美元。**亟需**子系统，构建“监测-分析-处置-优化”智能闭环，实现大模型应用的**全局可视、风险可管、处置可溯。**

2.2 面临挑战

大模型的应用给企业带来便捷的同时也带来了新的安全边界问题，例如大模型输入/输出的敏感信息泄漏、对大模型的注入攻击等，应对这些新问题企业会新增新型的大模型安全防护手段，但这些手段基本针对某一单一的风险点，对于企业管理者满对众多大模型的安全问题缺少集中安全问题管控的顶层协同机制。

1. 数据孤岛效应：

覆盖的私有API通道注入恶意指令时，MASB无法独立识别跨系统攻击链。



的隐蔽指令链）；

- **缺乏因果推理：**当模型输出泄露客户隐私时，现有工具无法追溯至具体训练数据源或违规访问行为。

3. 告警有效性无法保障：

- **误报率居高不下：**各子系统独立告警导致重复通知（如MAF和MAVAS同时报告同一模型的异常行为），SOC团队日均处理告警量超千条，有效告警识别率不足20%。
- **大模型风险漏报：**攻击常采用“低频慢速”策略（如每月一次模型参数篡改），分散的子系统监控难以捕捉此类长期潜伏威胁。

4. 性能与安全的平衡困境：

- **监控影响业务：**MAVAS深度扫描导致模型推理延迟增加30%，企业被迫在安全性与业务连续性间取舍；
- **资源浪费严重：**多套子系统独立运行，硬件资源利用率不足 40%，运维成本飙升。

2.2.2 缺少大模型资产使用的合规性持续监测评估

企业在多元大模型部署模式下，面临安全与合规管理的结构性难题：

1. 大模型合规评估监管不足：

- **缺少评估监测机制：**企业内部部署或引入第三方大模型时，可以因缺乏全面系统的安全评估机制，导致“带病上线”风险（如训练数据污染、隐私泄露漏洞）

■ **动态监管滞后**：监管部门对大模型输出的新兴风险（如深度伪造内容生成）持续

子系统。

更新要求，企业难以及时同步至所有

2. 资产不可见性：

GPT、Claude等公共模型处理客户数据（如

● **“影子AI”泛滥**：员工私自调用Cha

医疗诊断报告），导致敏感数据通过非受控渠道外流。据Forrester

金融交易记录、医

调查，67%的企业无法完整识别内部大模型资产。

● **供应链风险隐患**：外部模型（如通义千问、DeepSeek）的数据存储策略、训练集

来源缺乏透明性，可能违反《数据安全法》关于数据出境的要求。

3. 权责管理缺失：

■ **访问权限粗放**：MASB虽能控制模型访问权限，但无法基于数据敏感性动态调整（如

研发部门可访问通用模型，但禁止调用含客户隐私数据的业务模型）。

■ **审计链条断裂**：模型使用记录分散在 MAF 日志、MASB 策略库中，无法快速生成

符合 ISO 27001 标准的完整审计报告。

1. 响应机制缺乏统一协同：

数据风险、合规标准

访问控制节点进行一

● **单点响应的局限**：大模型的响应处置往往需要综合输出/输出

以及具体风险行为综合判定后进行决策执行，如果只在某一访

● **策略冲突风险**：各子系统独立处置可能导致某环节认为合规的策略但在另一环节

的策略管理中进行了阻断，因此缺少一套来源于综合风险判定指导多环节协同工

作的管控机制。

2. 闭环治理缺失：

- **修复验证空白：**传统处置仅完成风险遏制，未验证后续修复效果（如模型参数回滚后未重新评估漏洞状态），

复用。

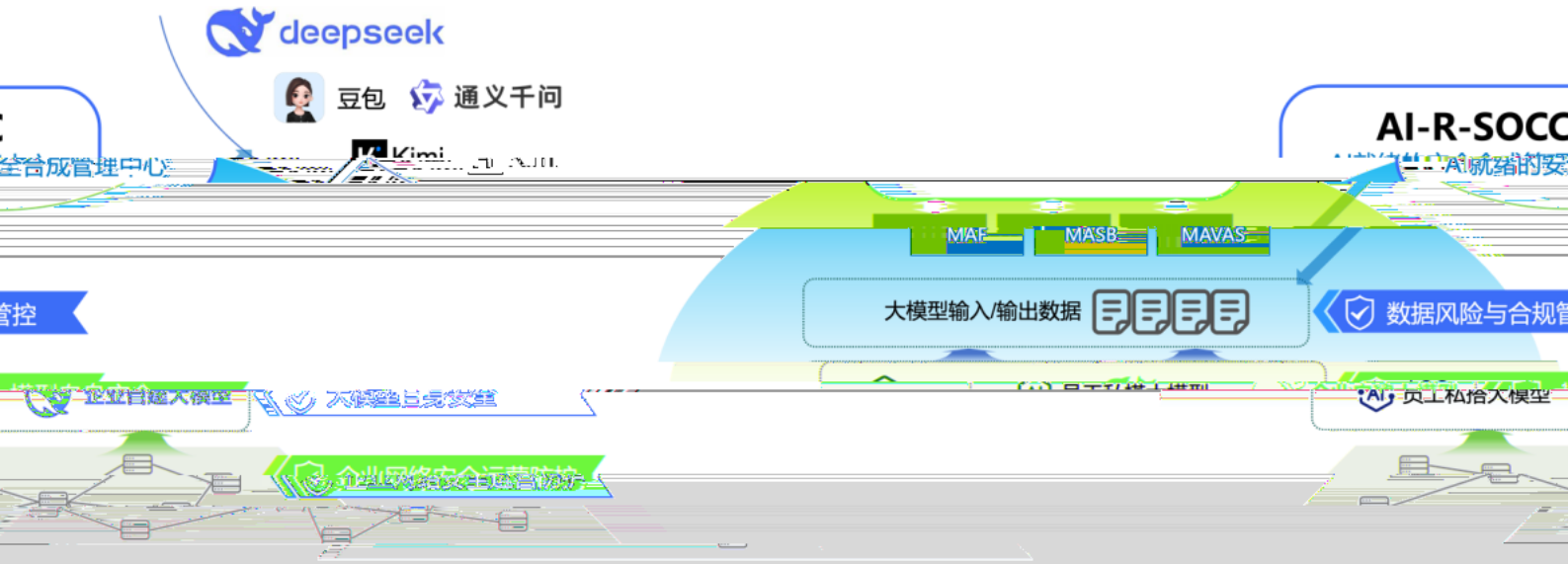
- **知识沉淀不足：**处置经验沉淀在人员头脑中，缺乏标准化剧本库供复

3 定义和建设思路

3.1 AI-R-SOCC 定义

面对大模型安全挑战，亟需三位一体的 AI 安全管理基座融合大模型应用安全、数据安

全融合管理中心（AI-R-SOCC）应运而生。



在企业员工大量使用大模型辅助办公的趋势下，大模型也成为了企业中的信息交互中心，这势必带来了新的安全边界问题，给企业安全运营工作带来了新的挑战。AI-R-SOCC 依托于完善强大的智能化安全运营支撑能力，同时应对 AI 安全新场景，通过纳管并有机融合 MAF、MASB、MAVAS 等大模型安全的管控手段，可形成对大模型应用自身的安全性评估、规性管控，并且这些过结合人员身份验证对大模型使用中所有输入/输出数据的风险性和合营防护，形成大模型应用程会被 AI-R-SOCC 放之于企业网络安全的整体视角中进行统一运营安全、数据安全、网络安全的一体化安全建设。

3.2 建设思路

中枢基

阻断大

全生命

AI-R-SOCC 是企业级大模型应用的集中化安全治理以及智能化安全运营保障的

座，在保障大模型使用安全的场景中可以统一纳管合规审核、身份管理、实时监测和

模型安全子系统（如 MAF、MASB、MAVAS）构建覆盖“准入-运行-处置-审计”全

身安全以及输入输出内容合规性的监测评估，通过 MAF+MASB+MAVAS 对

的大模型使用行为的风险性、合规性进行全面分析，发现威胁行为以及可能造

续的大模型自身

企业内部人员的

成的企业损失。

一协同治理的过程中实现大模型应用的全域可知、风险可防、处置可溯。基座

在这一统一

舌：

的核心定位包括

管理平台：作为企业大模型安全体系的“指挥舱”，聚合分散的安全能力，打

据孤岛与策略壁垒；

- **上层管**

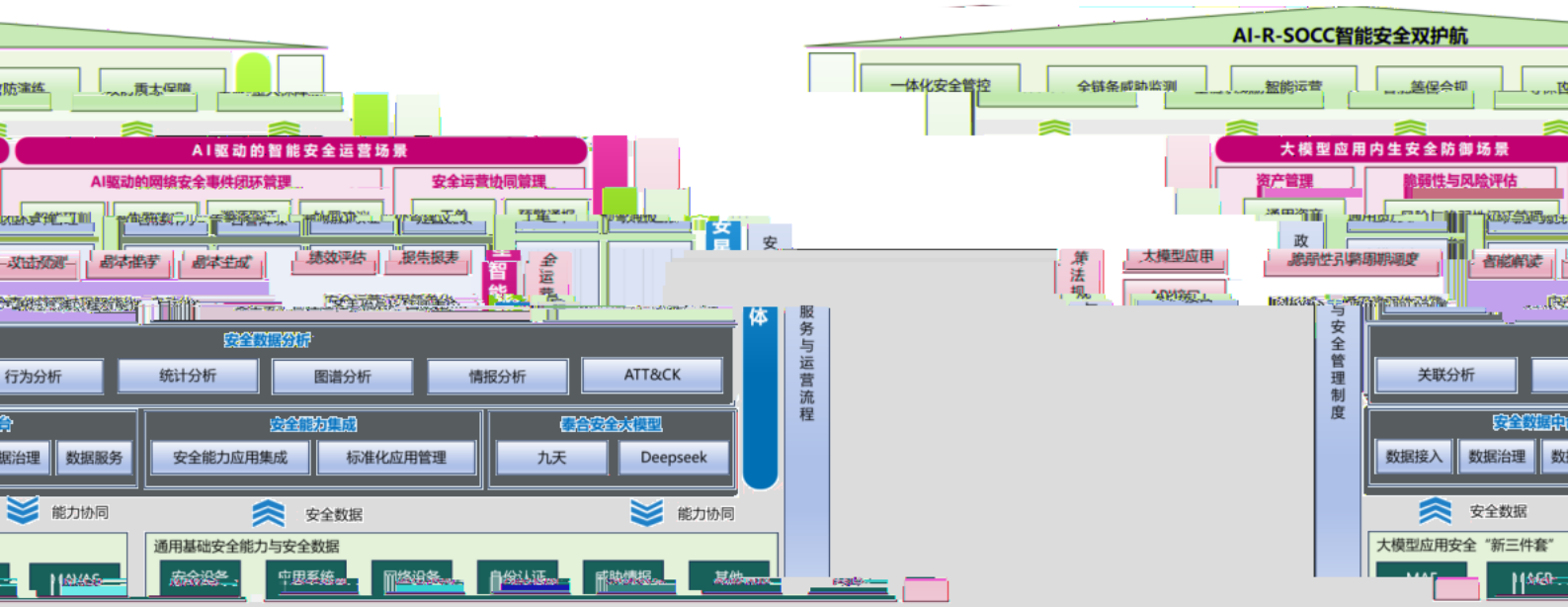
通数据

命周期治理：从模型上线前的安全合规审查，到运行期的实时监测与动态阻断，

- **全生命**

形成闭

闭环管理：



源异构的海量数据进行接入和治理，可接入融合包括大模型应用安全“新三

模型应用防火墙、MASB 大模型访问安全代理、MAVAS 大模型安全评估

类大模型安全和安全运营监测数据，同时平台提供了一系列数据集成和智能

掌上层能力构建，包括处置海量多源异构信息的数据中台能力、提供安星智

智能决策的泰合安全大模型能力，以及多种关联分析引擎、行为分析引擎、

平台提供多

件套”（MAF 大

系统）在内的各类

化分析引擎来支

能体辅助驾驶及智

图谱分析引擎。

AI-R-SOCC 提供了面向大模型使用生命周期的安全管控能力以及 AI 驱动

营能力。面向大模型的安全管控能力包括融合 MAF、MASB 和 MAVAS 的

性及合规性的综合评估能力、基于综合分析的大模型使用风险监测及安全事

风险行为的策略联动阻断能力。智能化安全运营能力包括通过 AI 驱动的智

在此基础上

的智能化安全运营

大模型资产安全性

件管理能力以及风

术提升安全运营形成智能安全双护航的全新解决方案。

及通过 AI 技术

4 核心能力

4.1 安星智能体

[illegible] AI-R-8000是专注于智能化处置一体

互，安星智 全事件协同处置、安全指令下发、剧本执行于一体的交互平台。通过自然语言交

全剧本，大 能体使安全人员能够在统一的界面上高效完成指令传递和执行，同时调用预设安

幅提升安全事件处理效率。

[illegible] 可精准解析 安星智能体支持7×24小时不间断响应，具备文件识别和文字理解等能力，

，帮助安全团队在 和执行安全指令。结合历史事件和知识库，智能体还能自动推荐处置剧本

复杂场景中快速制定精准应对策略。

，系统内置的小模 用户可以通过群聊@安星智能体唤醒其自动回复功能。对于简单任务

[illegible] 对话或鼠标操作完成动 反馈。用户还可以通过点击悬浮图标进入智能体界面，利用自然语言对

与快速响应。 作调用、剧本执行和文件上传等任务，实现对安全场景的全方位支持

4.1.1 能力市场

安全工具与协作工具标 能力市场为安全运营提供了一个整合工具与资源的平台，将各种

框架，开发人员可以通 准化封装为安全能力，并以服务形式呈现。系统具备开放的应用集成

础设施转化为可操作 过 Python、Java 等语言，借助内置 SDK 开发并集成应用，将安全

大模型的调度能力实现 的能力。每个应用包含一个或多个动作作为执行的最小单元，并通过

统一管理。

和纳管设备的实时监控，智能识别设备的连通性、性能

发现问题并采取防范措施，提升整体安全运营的效率。

杂场景下的自动化响应流程。

同时，系统提供对剧本执行情况

状态及潜在安全隐患，帮助用户提前

4.1.2 任务管理

各类任务能够得到及时、有效的处理。

大类型，确保安全运营中的

- 人工任务管理

办任务详情，用户可以快速查看、分配和处理各项任务，确保任

提升任务的执行效率。

通过直观的界面展示待

务按照预定计划顺利完成，

- 周期任务管理

功能，支持用户根据需求设置任务执行周期，并配置具体的执行

常定期检查与维护任务，减少人工干预，提升任务执行的自动化

与运维效率。

提供自动化的任务调度

动作，帮助用户高效管理日

水平与

- 脆弱性任务管理

专注于系统脆弱性管理，利用大模型的调度能力驱动漏洞扫描设备完成包括漏洞扫描、

令扫描、Web 扫描和配置核查等扫描任务，并对多次扫描数据进行对比分析，帮助用

速识别和修复系统中的安全漏洞，全面保障资产安全。

弱口令

户快速

4.1.3 大模型安全风险智能响应

智能响应围绕脆弱性监测这一核心任务，依托智能监测、智能值守和研判任务三大功能，构建了高效的安全监控体系。系统通过实时分析和动态跟踪，协同实现对脆弱性、告警事件的全面覆盖，有效提升威胁监测的精度与自动化水平，帮助用户实时掌控资产与业务系统的潜在风险，为安全态势感知与持续优化提供强大支持。

- 智能监测

通过大模型强大的实时分析能力，对系统脆弱性进行精准监控，支持用户创建自定义监

识别资产和业务系统中的潜在脆弱性风险，确保安全隐患

测任务，并结合最新漏洞情报快速

得到及时发现与处理。

- 智能值守

与任务跟踪能力，通过趋势图与列表展示脆弱性监测任务

务进展与资产状态。系统结合大模型的智能分析能力，将

行可视化呈现，为用户持续监测与动态分析提供数据支撑，

时跟进。

围绕脆弱性管理提供实时监控

的执行情况，帮助用户高效掌控任

任务执行效果与关联脆弱性资产进

确保脆弱性问题得到全面覆盖与及

- 研判任务

，帮助用户

事件响应效

警有效性判断、全路径分析以及画像分析，全面评估告警的威胁级别与潜在风险，

快速识别和优先处理高风险事件，聚焦关键威胁，减少不必要的干扰，显著提升事

率。

- 大模型风险处置

预置联动剧本，覆盖大模型应用安全、恶意使用、违规输出等场景，例如：

- **提示词攻击处置**：MAF 检测到恶意指令注入 → AI-R-SOCC 触发 MASB 冻结账号

输出内容 → MAVAS

- **数据泄露阻断**：MASB 识别敏感文件上传 → MAF 实时脱敏

据治理部门。

扫描关联模型训练集残留风险 → 隔离高风险模型并告警数

误封合法用户）。

理

4.2 全局资产管

管理管理能力，帮助企业“摸清家底”，通过多种适配器的有机融

平台提供全面的资产

模型应用场景下，新增了针对大模型的资产管控，使组织能够从全局掌控大模型应用相关的资产状况，为风险管控、全局监测提供基础支撑。

4.2.1 大模型资产实体定义

大模型资产包括大模型本身以及创建训练大模型、承载大模型使用过程中所涉及的各类资源和组件。这类实体资产主要包括：

- **大模型实体**：经过训练的大型模型文件，包含模型的架构、参数和权重等信息。
- **基础设施**：大模型应用部署的服务器、云主机、算力服务器（GPU），以及关联的网络资源，提供完善的管理能力。
- **应用软件**：对大模型应用相关的应用、组件进行统一管理，包括数据库、数据湖、应用软件等。

- **API 服务管理：**对大模型应用对外提供的服务进行全面的管理，并支持对 API 访问的行为进行全程审计。

4.2.2 企业/单位自建大模型

[illegible]

提供基本信息和安全信息管理，便于对此类大模型进行全生命周期治理。

[illegible]

4.3 大模型安全分析

引擎、异常行为分析引擎、

智能降噪、事件自动溯源等

平台基于多源异构的全量安全数据，通过分布式关联分析引

UEBA 画像分析、ATT&CK 分析、知识图谱分析，并结合告警

技术横向对传统安全威胁以及大模型应用场景的安全威胁进行实时监测，智能分析一个或

人员行为”的全维度安全分析体系，实现从单点检测到全局洞察的跃升。

4.3.1 大模型风险关联分析

的多维数据

安全运营场

手段，支撑安全

平台为组织提供高性能、高可靠、可扩展的分布式关联分析引擎，通过实时

关联能力，将 MAF、MSBA 以及其他类型的安全数据进行实时分析，不仅覆盖

景的数据关联分析需求，也为大模型应用安全场景提供了灵活、高效的分析手

场景的有效落地。

件的技

通过过滤、聚合、统计、关联等手段，发掘隐藏在这些数据之后的真实攻击或异常事

术，它可以辅助识别网络威胁和复杂的攻击样式，生成高级层面的安全场景。

用场景

平台接入 MAVAS、MAF、MASB 的安全数据、IAM、流量数据等对大模型应

度风险交叉验证，动态关联模型漏洞（如 MAVAS

如 MAF 捕获的提示词注入攻击），构建量化风

金融风控模型存在训练集残留敏感数据时，系统

记录，结合 MASB 权限日志中的越权访问行为，

的安全威胁进行实时检测与发现，实现多维

检测的 CVE 高危漏洞）、异常使用行为（如

险评估矩阵。例如，当 MAVAS 扫描发现某

自动关联 MAF 日志中该模型被高频调用的

综合判定风险等级并生成处置建议。

4.3.2 基于用户身份的行为分析

体（如主机、应用、网络流量和数据集）基于历史
线来进行分析，并将那些异于标准基线的行为标注
模型的打包分析来帮助发现威胁和潜藏的安全事

体的行为进行收集、处理和分析，主要关注网络中
过对网络中的行为深度解析和识别，将行为数据
建行为分析模型。

基于 MASB 同步的细粒度身份信息（包括用户角色、
C 构建动态用户画像，实现“身份-行为-数据”三位一
过关联 MAF 的输入输出审计日志、MASB 的访问控制记录及 MAVAS 的合规检
系统可精准识别高风险行为模式，分析能力例如：

- **风险行为评估**：整合各大模型安全子系统的数据，对用户的大模型使用行为进行风险和合规性分析。
- **行为画像**：建立用户安全使用画像，识别高风险用户、高频风险行为、关键违规场景和高频泄漏风险。

行为分析通过对用户和实体
轨迹或对照组建立行为轮廓基线
为可疑行为，最终通过各种异常
件。

行为管理负责对用户和实体
的行为模型和异常行为检测。通
和特征数据提供给模型管理，构

在大模型应用安全场景中，平台
部门归属、权限等级），AI-R-SOC
体分析。通
测结果，系

4.3.3 大模型安全图谱分析

监测管控过程中的所有主体、客
进行“实体—关系”的构建，基于将风险同实体关系模型映射后形成了直
行使用和自身风险全景视图，并通过关联展开或下钻的方式便于用户从总
模型安全风险状况。

AI-R-SOCC 基于知识图谱技术，将大模型安全生命周期
体、行为关系元素
观可视的大模型运
体到局部的掌控大

- 图谱实体涵盖：
- **风险实体：**用户（身份/权限）、大模型、业务资产、敏感数据信息元素、安全告警/事件、合规策略等。
- **风险关系：**使用关系、会话关联、输入关系、输出关系、事件归属、策略违规映射等。
- 基于知识图谱应用价值：

企业管理和运营人员可在构建的知识图谱中掌握当前的所有大模型风险、包括大模型的自身脆弱性及合规性评估结果、具有大模型不安全操作行为情况的员工、大模型使用中产生全的大模型应用为，便于在直观的“大模型风险地图”中不断的探索，外置最终得到一个安环境。

4.4 智能告警降噪

4.4.1 安全事件

多类告警聚合通过自动化方式生成(后分析自动化、精确化。

据组织，将关联的安全告警以事件级别、全告警的重要信息，同时在安全事件的全事件的影响范围。

式展示安全事件的攻击关系，方便用户关联的告警为边，并支持数据的下钻，

安全事件是告警之上的进一步数据凝练，由一类、续版本增加手工方式)，数量少、质量高，促进事件分

1）“一眼”看清安全事件来龙去脉：全视角数据关键路径、ATT&CK 等方式进行呈现，一眼看出安全概览页面将 IP、资产等信息进行展示，清晰展示安全

2）“一图”展示安全事件攻击故事线：以拓扑形的安全事件进行研判。拓扑图以 IP/资产为节点，以

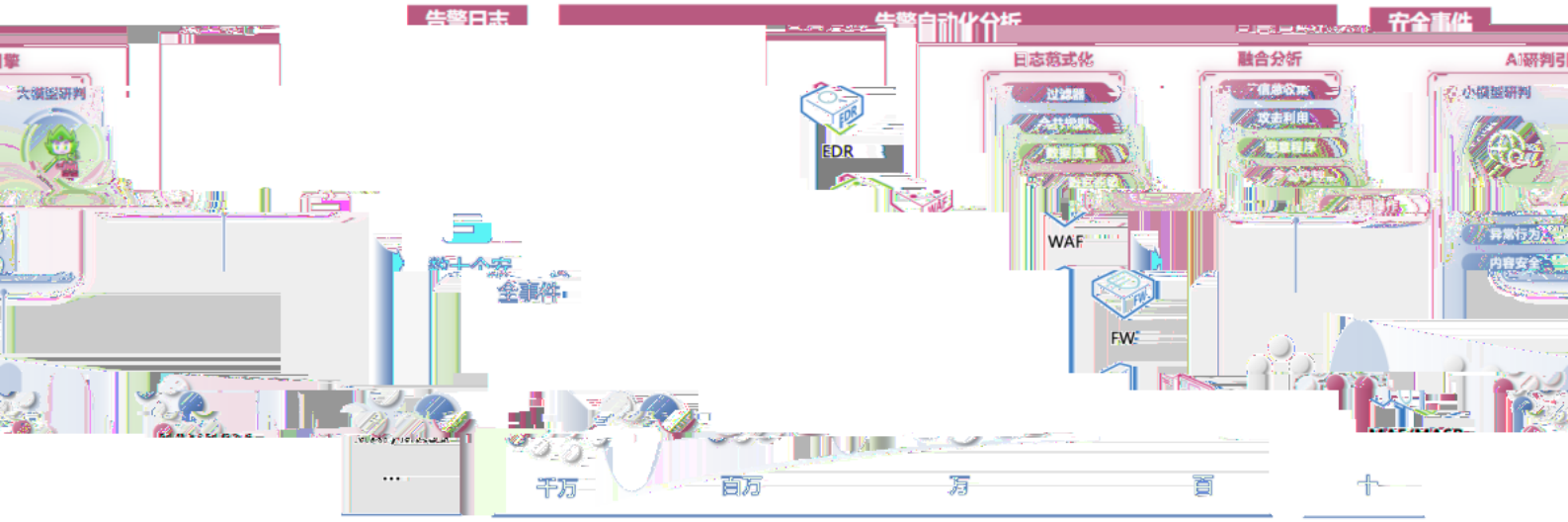
可以查看主机的资产信息、告警、异常行为、以及进程信息（需要接入启明星辰 EDR 日志数据）。

实体页签对关联的 IP、域名、账号、进程、文件等实体进行统一展示，支持调用安全设备、系统、应用进行相关的处置动作下发，实现对安全事件的处置、抑制。

4.4.2 智能降噪

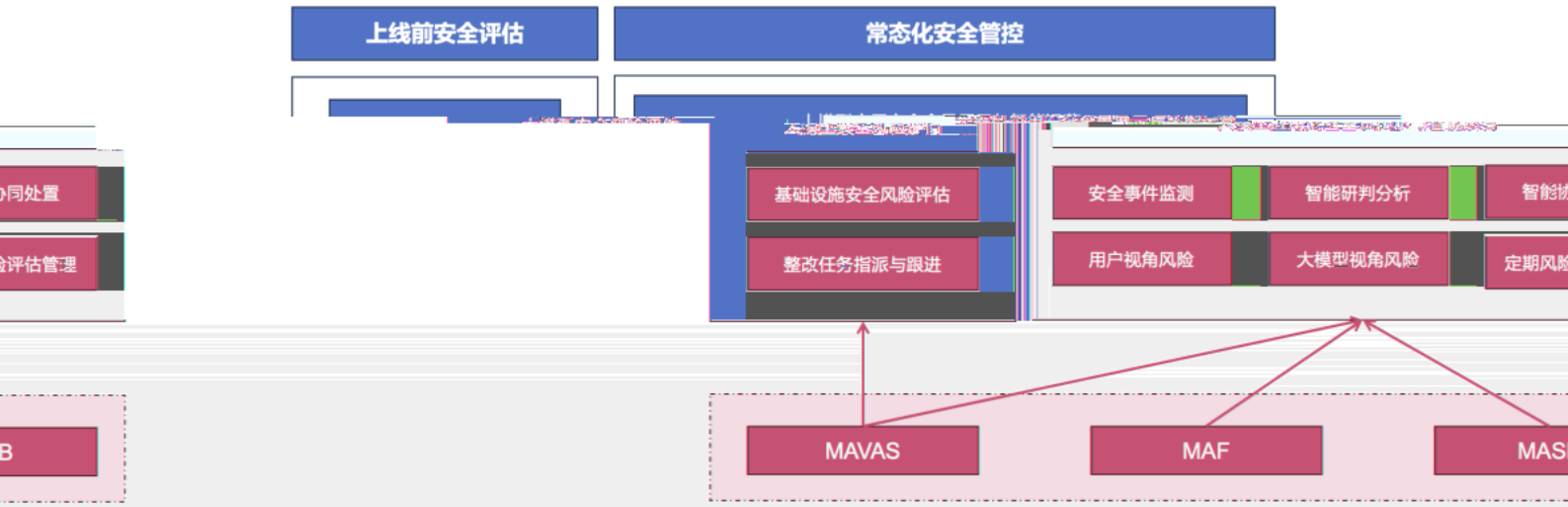
告警疲劳中解脱出来，投入到安全运营更关键的任务中。

全息降噪引擎将安全专家的知识与 AI 能力进行深度融合，覆盖常见的安全事件类型的智能研判与降噪，适用于大模型应用安全场景。



4.5 大模型风险监测管控

AI-R-SOCC 通过全景式风险感知技术，构建覆盖“管理个体-使用行为-业务环境”的多维监测体系，实现从风险发现的个点可见，升级为形成人机协同研判的闭环管理，助力企业实现大模型应用的风险可视化、处置精准化、治理智能化。



4.5.1 大模型自身风险监测

线前提供相应的

应用在不同应用

1) **上线前风险评估：**平台与 MAVAS 进行深度能力协同，在大模型上线

的风险评估。即通过大模型生成各种对抗攻击样本用于评估大模型应

2)

再到企业单位的大模型使用政策及规范

3) **内容合规实时检测**：动态维护国家到行业

（如《生成式人工智能服务管理暂行办法》等），对大模型使用过程中的使用和输

出数据进行实时监测，动态评估大模型的合规水平，便于及时整改

服务

4) **服务能力监控**：监测大模型的响应延迟、资源利用率，可与企业内多大模型的

接口联动进行服务自动触发熔断或负载均衡策略。

件的

5) **系统层风险预警**：扫描模型依赖环境风险（如部署服务器的系统漏洞或相关组

暴露面风险），生成支撑系统的脆弱性影响报告。

4.5.2 风险人员监测

时将监测到的大模型风险行为对应到实际使用人员上，同时可进行长时间周期的用户行为审计记录，实现大模型风险责任到人且问题事件可追溯，有效提升企业中应用大模型的风险可控。

2) **高危用户画像**：基于人员行为风险数据对企业中应用大模型的所有人员构建用户风险画像，标注典型的风险行为特征（如“生成违规内容”、“敏感数据违规输入上传”、“尝试注入攻击”等）。提供可视化视图呈现“高危人员榜单”、“风险人员分布”、“风险行为分布”等，便于由整体到个人掌握人员风险情况。

4.5.3 风险行为/事件监测

1) **多维度告警聚合**：将 MAF、MASA 所监测的大模型攻击行为、风险使用行为、恶意指令、越权访问记录等的告警事件，MAVAS 的合规性、风险性评估，AI-R-SOCC

查询和详情呈现。

的原始日志、行为分析甚至流量数据，可根据安全运营的诉求对告警事件进行深入

入排查并减少大模型的应用风险。

4.5.4 智能治理建议生成

AI-R-SOCC 基于安星智能体可智能化的对监测的威胁行为活风险隐患快速精准的生成处理建议，并综合企业的大模型应用环境给出治理建议，帮助用户推进大模型的安全合规化使用。

4.5.5 大模型风险管控处置

AI-R-SOCC 可基于安星智能体提供安全风险的联动响应，实现安全监测到处置的闭环，具体可见“4.1 安星智能体”章节。

4.6 行为审计溯源

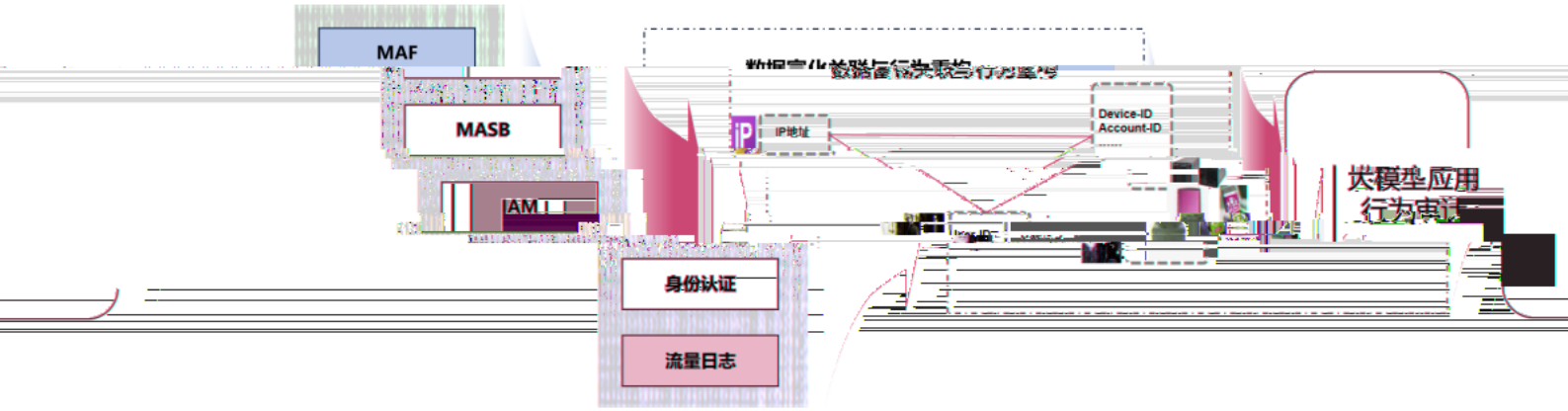
平台支持对大模型应用的访问行为进行全程监测与审计，行为可追踪、事件可追溯，满

SB、MAF 的行为数据于告警数据，结合 IAM、身份认证等安全数据，将

平台对接 MAS

大模型应用的访问行为进行全程记录。

用户、业务协同对



4.7 大模型安全态势呈现

SOCC

则、风

行呈

在大模型多种安全风险评估、监测、管控的过程中将产生大量的监测数据，AI-R-S

通过**多源数据融合引擎**与 **2D+3D 可视化技术的应用**，将来海量的合规探测、行为监

险告警等数据进行整合聚焦形成大模型安全的态势大屏面向用户的管理及运营人员进

大模型资产风险态势、大模型使用风险监测

屏。

安全态势的呈现，综合所有的风险因素呈现

、使用中的安全威胁、用户行为安全的高纬

势。

全生命周期安全态势的呈现，覆盖内部部署

实体，整合漏洞评估、合规检测、供应链风

险。

AI-R-SOCC 提供了大模型综合风险态势、大

态势、大模型智能运营态势四个维度的态势大屏

- **大模型综合风险态势：**面向大模型全局

  总体安全健康指数以及大模型自身风险

  指标，便于全局掌控安全风险和安全态
- **大模型资产风险态势：**面向大模型资产

  模型、私搭模型及外部公共模型等资产

与数据流动风险的实时监

- **大模型使用风险监控态势：**面向大模型使用过程中人员行为

权限滥用、敏感信息泄漏、异常操作等威胁维度，通过多源行为日志

控，聚焦用户

于快速识别高危人员并阻断违规行为。

融合分析，便

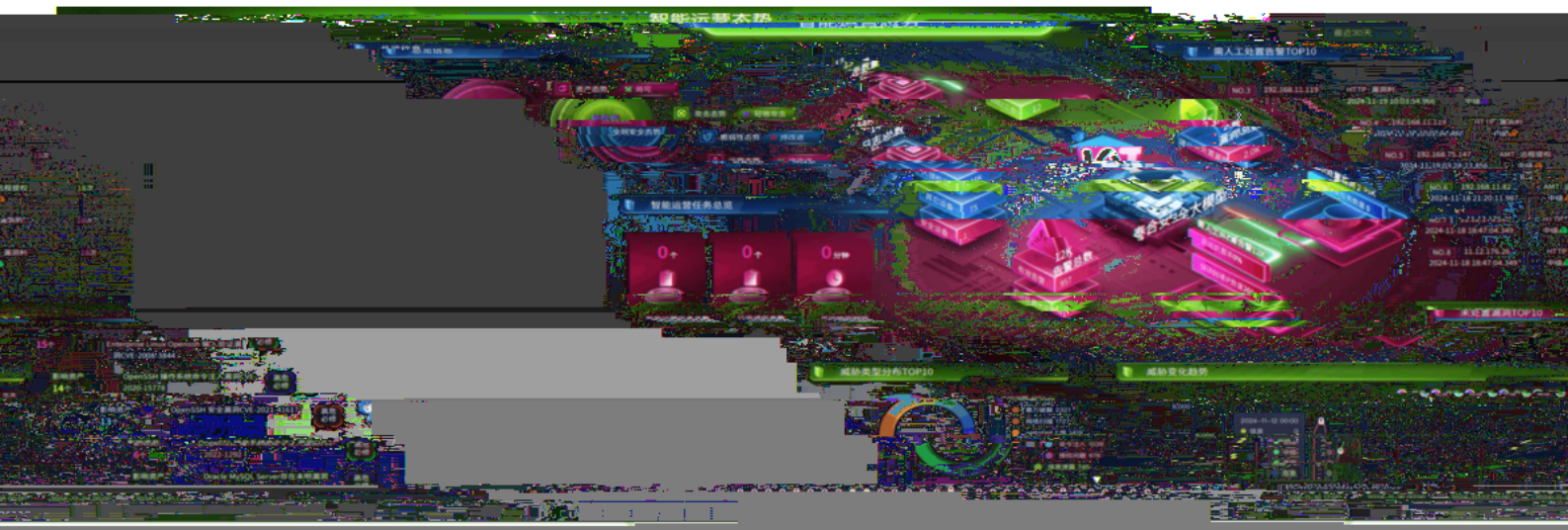
营态势：智能运营态势大屏集成了总览信息、智能运营任务完成情况、

- **大模型智能运**

TOP10、威胁变化趋势、需人工处置告警TOP10和未处置漏洞TOP10

威胁类型分布

有效优化资源分配与运营流程，为快速响应和精准防护提供了坚实基础。



用场景

5 部署和典型应

应用安全方案的核心基座，对接大模型应用安全“新三件套”

ASB 大模型访问安全代理、MAVAS 大模型安全评估系统），

AI-R-SOCC 作为大模型应

（MAF 大模型应用防火墙、M

杂安全需求，提供针对性和高效

全运营，针对大模型在训练、推理、部署等全生命周期的复杂

：

的解决方案，保障大模型安全稳定运行，如下为部署示意图

中实时风险管控

5.1 场景一：大模型应用

- **痛点**

1. **敏感数据泄漏风险**：用户输入或模型输出中隐含客户隐私（身份证号、银行卡号）、商业机密（设计图、定价策略）等敏感数据。
2. **生成内容滥用风险**：恶意用户通过诱导指令生成虚假信息（伪造财报）、违法内容（钓鱼邮件模板）、对抗性输出（绕过风控规则的话术）；

或篡改模型参数，导致输出结果偏差其

3. **大模型投毒攻击风险**：攻击者污染训练数据

至业务决策失效；

不符合《生成式人工智能服务

发法律处罚与品牌危机等。

4. **输出内容合规风险**：模型输出违反行业监管要求（如不

管理暂行办法》）、伦理准则（如歧视性言论），引发

- **解决方案**

1. 敏感数据泄漏防控

传文件与指令中的敏感字段，

✓ 通过 MASB 的 AI-Sensitive DLP 引擎，实时扫描上传

动态拦截高风险内容；

、商业秘密）、违法违规内

✓ 实时监测输出内容，识别敏感信息泄露（如客户住址

及时抑制风险的产生。

2. 生成内容滥用阻断

生成终过审 ✓ 语义深度解析：基于多轮对话上下文理解，识别隐蔽攻击指令（如“生

核的贷款申请话术”）；

✓ 合规规则库：内置《大模型系统安全保护要求》等国家、行业相关大模型管理要

求合规策略（如禁止生成“保本保收益”承诺），实时拦截违规内容；

低。

降

模型投毒攻击防御

3. 大

入数据验毒：通过 MAF 检测输入的异常数据（如噪声注入），联动 MASB 阻

✓ 输

污染源 IP；

断

模型训练或投喂数据批量认证：对模型训练集进行 MAVAS 安全扫描，拦截未

✓ 大

认证数据源。

4. 输出内容合规治理

✓ 大模型对话敏感数据防泄漏：通过对输出数据的全面解析，识别大模型对话交互过程中的敏感信息输出实时识别、封堵或脱敏，保障大模型服务合规。

✓ 动态合规监测：输出内容实时比对监管要求（如《生成式人工智能服务管理暂行办法》），对不合规内容生成告警；

- **价值成果**

1. **数据泄漏防控：**如某政务平台实现公民隐私泄露事件归零，金融客户年均减少损失大幅降低；

低；

3. **减少业务损失：**避免因模型滥用导致的品牌声誉损害与用户流失。

5.2 场景二：模型上线准入管控

- **痛点**

"带病上线"风险（如训练数据污染、隐私泄露漏洞）；

2. 合规审查较多依赖人工，效率较低，平均耗时 2-3 周，影响业务创新节奏。

- **解决方案**

1. **自动化安全审查：**

[illegible]

目标劫持 测试库自动化评估模型在道德伦理、歧视偏见、财产隐私、反面诱导

评估模型抗攻击能力。 等 14 类安全隐患，量化

[illegible]

/API 敏感数据泄漏、大模型应用层漏洞攻击等方面进行充分实战验证。

✓ 内置行业合规知识库（如《大模型系统安全保护要求》、《生成式人工智能服务

模板是否符合监管要求。 管理暂行办法》），自动校验模型输出

2. **动态准入决策**：

阈值（如安全分<80、合规分<90）的模型 ✓ 采用量化评分机制（0-100 分），未达

给出禁止接入生产环境结论；

✓ 生成可视化评估报告，明确整改建议（如“训练数据脱敏率需提升至 98%”）。

- **价值点**

降低大模型上线风险，审查周期有效缩减。

[illegible]

- **痛点**

orFlow）存在未修 ✓ 模型漏洞暴露风险：大模型依赖的软件框架（如PyTorch、Tens

复的CVE漏洞（如CUDA内存泄漏漏洞），可能被攻击者利用进

行模型逆向攻击或 参数篡改。

✓ 合规状态失守风险：模型输出内容因监管政策动态更新或训练数

据缺陷（如未脱敏 个人信息）导致合规偏离。

PU利 ✓ 服务能力退化风险：模型API响应延迟飙升（P95>1000ms）、资源过载（G

用率>95%）引发业务中断，且缺乏实时预警与自愈机制。

1. **漏洞实时监测与修复：**

✓ 漏洞扫描：联动 MAVAS 每小时扫描模型依赖环境（如 CUDA 版本、容器镜像），识别高危漏洞（如 CVE-2023-1234）；

✓ 自动化补丁：对低风险漏洞自动推送修复建议（如升级 TensorFlow 至 2.12），高风险漏洞触发模型隔离并告警。

2. **合规性动态适配：**

✓ 策略同步引擎：实时对接监管机构数据库（如网信办的合规库），自动解析最新条款并转化为检测规则（如禁止生成“深度伪造视频”）；

✓ 输出内容校验：通过 NLP 引擎比对模型输出与合规知识库，违规内容实时脱敏或阻断。

保障：

3. **服务能力**

线监控：设定 API 响应延迟（<500ms）、错误率（<1%）、资源利用率

✓ 性能基线

PU 节点或切换

✓ 智能自愈：当检测到资源过载时，自动基于设定的剧本策略扩容 G至轻量化模型版本。

- **价值成果**

1. 拦截多起利用 PyTorch 漏洞的模型逆向攻击，避免客户数据泄露。
2. 政务问答模型输出内容合规率提升至 100%。

5.4 场景四：影子大模型监测与治理

- **痛点**

  ✓ 企业大模型资产清单分散在多个部门，存在未登记的“影子模型”；

  ✓ 员工私搭模型（如 Llama 2 代码生成工具）成为数据泄露与攻击跳板

- **解决方案**

  ✓ 自动发现企业内部模型实例（包括容器化部署的私搭模型），构建动态资产库；

  ✓ 通过流量探针（识别非标准 API 特征）与代码扫描（检测私有模型仓库），自动构建私搭模型指纹库。

  ✓ 通过对发现的影子大模型进行风险探测和输入/输出监测，对风险进行提示并在必要时实时阻断。

- **价值成果**

  • **大模型资产黑洞消除**：识别并治理若干个未登记的私搭模型，收缩企业内大模型攻击面风险。

6 能力优势

6.1 智能运营

基于大模型安全的深度洞察，智能运营模块精准适配大模型安全运营场景。通过对大模型训练数据来源监测、推理过程行为分析等任务，实现大模型安全运营的日常工作自动化处理。利用先进的可视化技术，让数据流向、风险检测点、任务执行进度等关键信息清晰直观，使工作进展透明可见，帮助用户快速定位问题环节，高效跟踪任务闭环，确保大模型在安全的环境下持续运行。

6.2 集中调度

充分考量大模型安全涉及的多维度安全能力，将各类针对大模型数据安全防护、模型[illegible]法保护、运行环境加固等安全平台能力统一管理调度。打破安全能力“烟囱式”隔离[illegible]

6.3 快速响应

[illegible]攻击时的紧急处理、数据泄露风险的应急响应等，能够快速、并行地进行处理，在极[illegible]内启动多套应急预案，提升安全运营响应效率，最大程度减少安全事件对大模型的影响。

6.4 安全专家

依托强大的大模型知识图谱与数据分析能力，提供针对大模型安全的常用安全知识智能

分析大模型运行过程中的安全数据，包括访问日志、异常行为记录等，为运

答。同时，深入

的安全态势分析报告，弥补运营人员在大模型安全领域的经验差距，提升整

营人员提供深度

容应对各类大模型安全威胁。

体作战实力，从

采用先进的自动化监控与智能预警技术，实现 7×24 小时全年无间断的大模型安全运

营值守，对大模型运行状态进行实时监测，无论是节假日还是深夜，一旦发现威胁大模型安