

TVP「AI与安全」高峰论坛

大模型时代 安全如何洗牌



AI赋能的未来安全 创新趋势和实践探索

主讲人

深信服 张振礼

01

核心认知

AI时代安全理念的变化
对AI赋能安全的理解

02

产品战略方向

AI时代安全产品规划思路
产品战略规划的方向

03

核心技术介绍

投入情况、训推能力
安全语料工程化和微调能力
基模S1的打造过程

04

关键场景实践

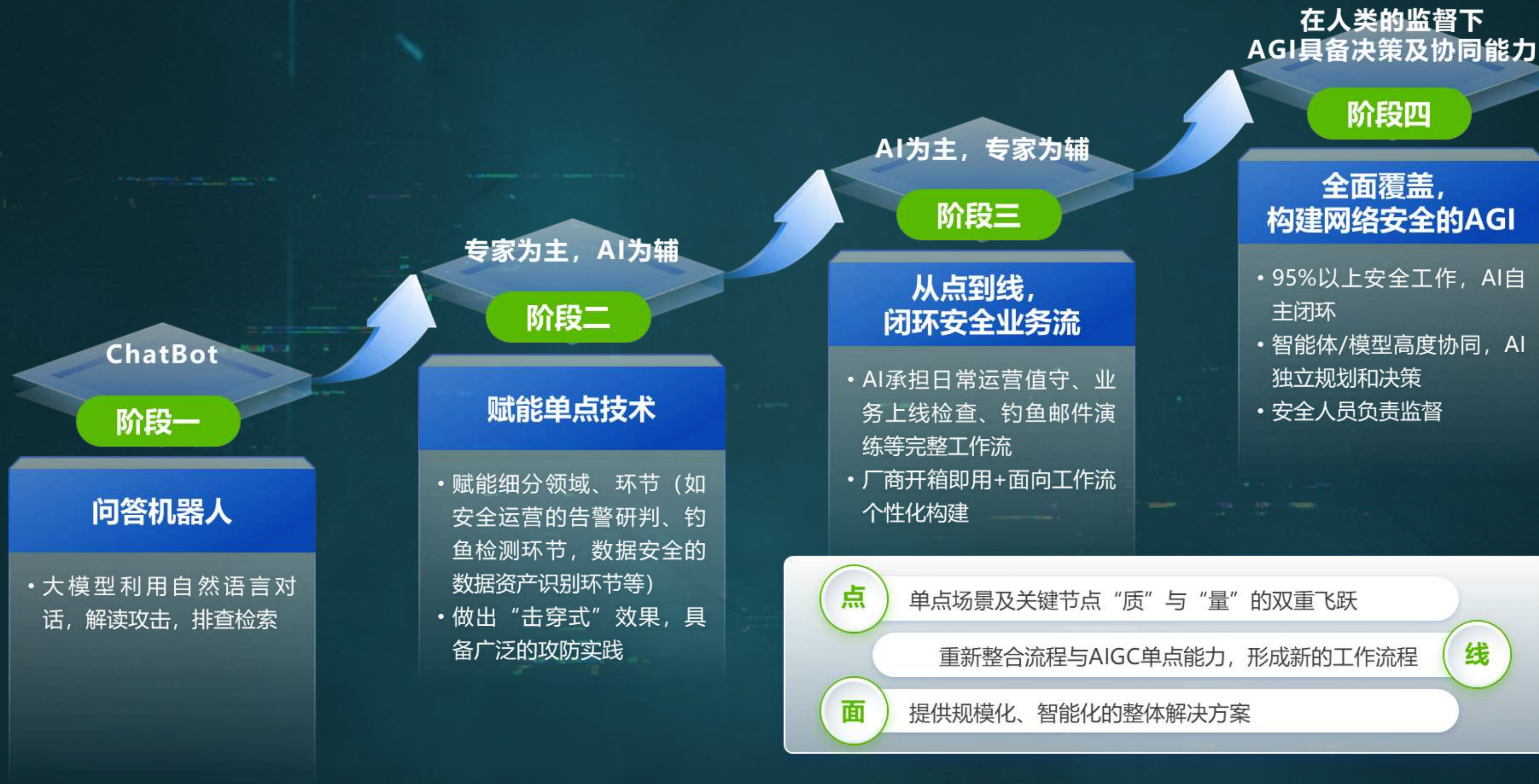
关键场景的选择和落地情况
客户和市场的反馈

01

核心认知



AI必将深刻改变网络安全工作





安全产业对大模型的追捧，从兴奋变为冷静

产业界对大模型的追捧，从兴奋变为冷静

2023

- 探索之年
- 大预言模型之年
- 颠覆和高光时刻

2024

- 执行和结果
- 多模态模型之年
- 失败、泡沫破裂

Gartner

从技术探索回归大模型的商业价值实现

75%的受访企业表示“实现真正的降本增效”是企业应用AIGC的首要目的

75%，降低成本和提升效率

36%，提升市场的反应速度

34%，产品差异化创新需求

数据来源：红杉中国2023年CIO调研问卷



在网络安全领域，用户对大模型应用也持有不同的态度

2023~2024，我们在推进安全大模型的过程中，听到了几种不同的声音

乐观的

- 大模型技术是未来，只要经过充分的训练和调优，安全智能化程度将迎来飞跃，能够解放安全运营人员
- 大模型技术会不断演进，应该紧跟趋势
-

怀疑的

- 主要看到的是聊天场景，没有解决实际的安全痛点
- 通用模型在真正的专业领域的错误和幻觉还是比较多
- 蹭热点，感觉是包装营销
-

观望的

- 技术是好的，但针对我的需求，解决到什么程度还不清楚
- 技术门槛相对过去高了许多，而且收入产出比可能不划算，再看看
-

业界对大模型的态度逐渐进入冷静期，2025开始需要回归到实际价值的实现



对AI赋能安全的理解

第一，安全本来就碎片化，一定不能每个场景单独的模型分开部署，要统一规划，通过良好的架构设计实现各类安全大模型算力的统一调度和适配

第二，基础大模型会持续进化，安全大模型需要具备快速融合、集成、吸收优秀基础模型的知识 and 推理能力，智能体的元年开启，持续为安全场景带来体验和效果的改进

第三，安全管理团队成员在AI时代也要具备RAG、微调等技能，面向全员的安全应用和主动审计智能体会彰显团队价值

第四，大模型时代会逐渐改变甲乙双方传统的建设与协同关系，从提供封闭产品到提供开放AI安全能力，通过提供灵活的开放性，匹配用户自主开发创新的诉求



安全垂域大模型的构建，是一个系统性工程

阶段一 构建安全基础 大模型

阶段二 构建细分领域 大模型

阶段三 构建细分领域 安全智能体

阶段四 部署上线交付 与数据飞轮

通用开源基模

deepseek

通义千问

ChatGLM

增强预训练

指令微调

深信服安全
基础大模型

蒸馏
cot强化学习
细分领域SFT
...

安全运营
大模型

流量威胁检测
大模型

钓鱼威胁检测
大模型

数据风险分析
大模型

行为风险分析
大模型

.....

模型推理优化

Prompt优化

智能体
开发与集成

安全运营
智能体

流量威胁检测
智能体

钓鱼威胁检测
智能体

数据风险分析
智能体

行为风险分析
智能体

.....

高素质本科生:

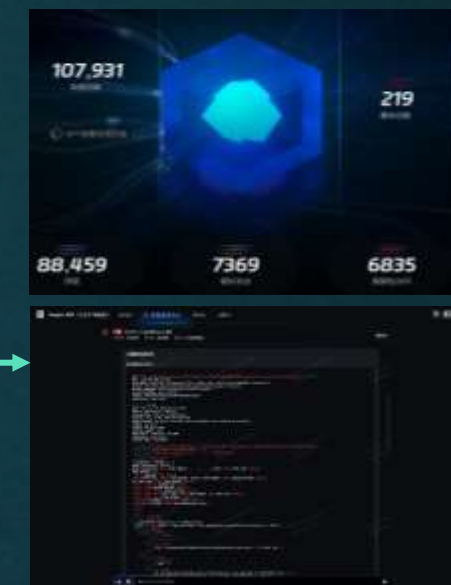
- 良好的安全知识和推理能力，安全领域领先通用基础模型
- 良好的对话和生成能力，但未经过细分领域优化，存在一定幻觉
- 可进一步微调，构建安全细分领域大模型

专业精深的研究生:

- 高质量细分领域指令微调，召回率和精准率显著提升
- 领域内幻觉问题得到良好控制

具有丰富经验的专家:

- 面向具体场景，搭建复杂系统工程，融合场景工具
- 大小模型、专家模型混合，提示词上下文关联，系统吞吐提升，合规兜底...
- 面向具体场景实战能力优秀





我们的投入

TVP Tencent Cloud
Valuable Professional

预训练模型

基于ChatGLM、LLaMa等多个优秀基础模型打造，MoE架构，参数总量近千亿

训练硬件资源

自有800+张A100/A800显卡，云上还有部分数据中心的支持，每轮训练~20天

推理运行要求

蒸馏、剪枝、量化技术压缩模型到百亿参数，8*4090运行运营大模型，RTX4090/4080运行检测大模型

二次预训练

3千亿token的高质量通用和安全领域语料（1TB+训练语料）

人力资源投入

400余名硕博士专职投入（创新研究院、网络安全能力部、终端安全能力部等）

国产化适配

逐步适配华为昇腾、天数智星

微调

800万通用和安全领域指令具备丰富上下文的
安全场景微调数据对≥300w

研究时间投入

2015年积累小模型，2022年底大力投入大模型，2023年5月V1.0，2023年9月V2.0，2024年1月V3.0

客户资源

400余家客户已接入安全GPT，深度使用客户已近1年，单客户日均问答上百次



优势:

- 分析质量更高，回答逻辑性强，具备多维度信息关联推理能力，能研判得更深
- 通用知识、安全知识、代码能力显著提升，相关知识点问答更准确
- 生成内容天然包含推理think过程，研判过程更透明，甚至具备启发性
- 强依赖模型推理能力的场景，可以快速实现能力提升，比如BES邮件检测

短板:

- 与经过针对性训练和工程优化的垂域模型相比，绝大多数场景检出率和准确率明显偏低
比如终端安全各类测试集检出率在50%-70%
- 专业领域的能力比较欠缺，比如用户侧的资产识别效果、复杂事件溯源、安全工具使用
- Think过程其实增加了算力开销，慢思考的过程实际落地要增加GPU资源，还会输出琐碎词语



融合DS之后，能够给出较为准确和详实的思考过程，可有效协助用户确认研判结论的准确性和有效性

ds_rl_qwen32B Today at 13:31

Thought for 几秒 ^

“好，我现在要分析这两起异常行为告警（UEBA），看看它们之间是否存在关联性，并判断这次事件属于真实威胁、业务行为还是无法研判。

首先，看看第一起告警：“Anomalous SMB to Regular Locations”。威胁类型是异常的SMB连接到常规位置。攻击技术包括横向移动和内部侦察。告警等级是4级，显著行为。源IP是10.1.18.67，目标IP是10.1.18.188和10.1.18.182，都是445端口。时间是2021年3月13日23:04到23:53。补充信息显示SMB连接数是13，属于不寻常的使用，方向是出站，源IP在办公区/打印机区域。说明这个设备在短时间内进行了大量SMB连接，可能是在尝试访问多个内部设备，这可能被攻击者用于横向移动。

第二起告警是“Anomalous SMB”，威胁类型是异常SMB连接，等级3级，中等影响。源IP同样是10.1.18.67，目标IP是10.1.18.127，445端口。时间在23:02:22，只发生了一次连接。补充信息显示是出站，应用协议是SMB，目标端口445，源IP在办公区/打印机区域。这次连接虽然只发生了一次，但同样涉及SMB协议，可能是在尝试连接内部设备。

接下来分析关联性：两个告警的源IP都是10.1.18.67，目标IP都是同一个子网10.1.18.0/24内的不同设备，且都使用了445端口，这是SMB的标准端口。时间上第一起告警持续了约50分钟，第二起在第一起开始前两分钟发生。补充信息都提到源IP在打印机区域，这可能意味着该设备是打印机，通常打印机不太会频繁进行SMB连接，尤其是到多个设备。

从攻击链阶段来看，第一起告警涉及横向移动和内部侦察，这表明攻击者可能在尝试控制更多的设备或收集内部信息。第二起告警虽然没有明确的攻击链阶段，但同样涉及SMB连接，可能是横向移动的一部分。

综合来看，两起告警都涉及同一台设备（10.1.18.67）在短时间内进行大量SMB连接，目标都是同一子网内的设备，且源IP属于办公区打印机。打印机通常不会有频繁的SMB活动，尤其是到多个设备，这可能表明该设备被入侵，正在被攻击者用于横向移动。因此，这可能是一个真实威胁，攻击者试图在内部网络中扩散。”

分析结论

最终结论：真实威胁

分析原因：

1. 关联性分析：

- 两起告警共享相同的源IP（10.1.18.67），表明同一设备在短时间内发起了多次异常SMB连接。
- 目标IP均位于同一子网（10.1.18.0/24），且使用445端口，进一步支持了横向移动的可能性。
- 第二起告警的时间早于第一起，可能是在尝试建立后续连接前的侦察活动。

2. 威胁特征：

- 源IP属于打印机区域，打印机通常不应频繁进行大量SMB连接，尤其是到多个内部设备。
- 第一起告警显示13次SMB连接，属于异常行为，可能涉及数据窃取或横向移动。
- 攻击技术涉及横向移动和内部侦察，符合典型的网络攻击手法。

面向内网SMB连接访问，进行深度关联推理

<think>

> 好的，我现在需要分析这封钓鱼邮件。首先，按照步骤检查邮件头部。发件人IP是38.240.44.218，来自未知的IP，虽然发件人显示为DHL，但域名date-dxb.online的注册时间只有1年，这可能有问题。SPF记录可能没有正确配置，但这里没有明确的验证结果，需要进一步检查。

>

> 接下来看邮件正文，主题是“货物到达通知”，通常DHL会用这样的主题来引起收件人注意。正文内容简短，要求查看附件，但附件是一个HTML文件，这不太正常，通常物流公司会用PDF或文档格式。附件名是“DHL# Shipping_Documents.html”，可能包含恶意内容。

>

> 然后分析附件。HTML文件里有一个登录表单，要求输入邮箱和密码，并有记住我的复选框。这里明显是伪造的DHL登录页面，目的是窃取用户的凭证。页面结构简单，但输入框预填了收件人的邮箱，这可能增加可信度，诱使用户输入密码。此外，另一个HTML部分有“DD”字样，可能是个错误或拼接的痕迹。

>

> 发件人域名注册时间短，且附件中的登录表单设计粗糙，不像正规DHL邮件，进一步确认这是钓鱼企图。没有提到其他链接或二维码，但附件本身已经足够危险。

>

> 总结异常点：邮件头部发件人IP和域名可疑，附件是HTML表单窃取凭证，域名注册时间短，页面设计异常。处置方法应包括标记为钓鱼邮件，删除附件，提醒用户更改密码，并检查是否有其他受影响账户。

</think>

逐步研判邮件各个元素，综合总结是否异常

事件综述：

经过我的研判，这是一封伪装成DHL物流通知的钓鱼邮件。邮件通过“货物到达通知”主题诱导用户查看附件，试图窃取收件人的邮箱账号和密码。攻击者伪造了DHL官方登录页面，要求用户输入企业邮箱凭证，若用户轻信提交将导致企业邮箱账户被盗，可能引发内部信息泄露、钓鱼邮件二次传播等安全风险。

该邮件附件包含恶意HTML文件，通过伪装成物流单据实现钓鱼攻击。攻击者特别预设了收件人邮箱地址作为表单默认值，增强了欺骗性。若用户输入密码，攻击者可利用窃取的凭证登录企业邮箱实施横向渗透，威胁企业数据安全。

异常点详情：

- *件头部异常*1. 可疑发件人域名：发件地址global.sources@date-dxb.online使用的域名注册仅1年（创建于2023年），与DHL官方域名（dhl.com）无关联，不符合企业长期运营特征
- *邮件附件异常*1. 恶意登录页面：附件“DHL# Shipping_Documents.html”包含伪造的DHL登录表单，预设邮箱aaron.shi@quantacn.com作为默认值，诱导用户输入密码
- 2. 异常页面结构：页面存在双重HTML框架（一个完整页面和内部空框架），包含“DD”乱码字符，显示开发过程中抄袭代码的痕迹

如何处置这个事件：

- 立即将邮件标记为钓鱼邮件，禁止转发或点击附件
- 删除附件“DHL# Shipping_Documents.html”并清空回收站，防止误触激活
- 通知所有员工修改邮箱密码，特别提醒aaron.shi@quantacn.com账户持有者检查登录记录
- 在邮件网关拦截包含date-dxb.online域名的邮件，阻断后续钓鱼攻击

✘ 高级钓鱼威胁检测场景中，DeepSeek更容易出现结果跳变和幻觉

在钓鱼检测困难场景的多样本测试中，同样的检测任务和Prompt，出现结果跳变和思维链输出偏差，幻觉控制不够稳定

好，我来分析一下这个邮件是否是钓鱼邮件。首先，邮件主题提到这是更正后的发票，并且发件人是孙悟空，来自Faiwu@ahhhhst.com。收件人是李经理，看起来内容是工作相关的，尤其是涉及到发票，这可能让收件人放松警惕。

接下来看发件人的域名ahhhhst.com，注册时间只有一年，这可能是一个新注册的域名，有时候钓鱼网站会使用新注册的域名来规避检测。此外，邮件中有一个附件，这是一个常见的钓鱼手法，附件可能包含恶意软件或者链接到钓鱼网站。

邮件来源标记为外部，系统提醒收件人确认发件人可信，这一点也是提示可能存在风险。邮件头信息显示SPF检查通过，但发件人HELO标识的SPF状态为None，这可能意味着发件人的身份验证不够严格，增加了被钓鱼的可能性。

综合来看，虽然邮件内容看起来像是正常的商务沟通，但发件人域名较新、附件的存在以及外部提醒都增加了风险因素。因此，我判断这是一封钓鱼邮件，置信度为80分。

是否是钓鱼邮件：是

置信度：80分

嗯，我现在要分析这个邮件是否是钓鱼邮件。首先，我得仔细看看邮件的内容和相关信息。

邮件的标题是“李经理，您好。这是更正后的发票，请查收。”看起来好像是公司财务部门发来的，内容涉及到发票，这可能是一个常见的钓鱼手法，因为很多钓鱼邮件都会以发票、账单等名义出现，试图诱骗收件人点击链接或下载附件。

发件人是孙悟空，邮箱是Faiwu@ahhhhst.com，看起来像是财务部的邮箱。但我要检查这个域名的注册信息。邮件里的whois信息显示，ahhhhst.com的注册年龄只有1年，而且注册邮箱是通过阿里云的whois页面，这可能有问题，因为很多钓鱼网站会选择新注册的域名，或者使用不容易追踪的注册邮箱。通常，可信的公司邮件域名会有更长的注册历史，或者使用公司自己的注册邮箱。

接下来，邮件中有附件，是一个PDF文件，文件名是“火星宇宙有限公司〔总金额：3373686元〕.pdf”。虽然看起来合理，但附件可能是恶意的，比如包含病毒或者钓鱼链接。不过，邮件中没有明显的链接，这点可能降低风险，但附件本身也可能有问题。

再看看邮件的SPF记录，邮件服务器显示SPF=Pass，说明发件人的域名通过了SPF验证，这可能意味着邮件确实是来自该域名。但是，SPF=Pass并不一定意味着邮件是安全的，因为攻击者可能已经入侵了该域名的服务器，或者使用了合法的域名进行钓鱼。

邮件的发件人IP地址是115.124.28.3，这个IP可能在阿里云的服务器上，因为邮件中提到postmaster@out28-3.mail.aliyun.com。这可能意味着邮件是通过阿里云的邮件服务发送的，但这也可能被钓鱼者利用，因为阿里云的服务可能被滥用。

总的来说，虽然邮件看起来有一定的可信度，但有几个因素让人怀疑。首先是域名注册时间较短，其次是发件人可能使用了云服务的邮件服务器，增加了隐藏真实身份的可能性。附件可能包含恶意内容，虽然没有链接，但下载附件也可能导致感染。因此，这封邮件有一定的风险是钓鱼邮件，但没有足够的证据可以确定，所以可能属于低风险。

根据分析，这封邮件存在一些可疑因素，但没有明确证据证明是钓鱼邮件。因此：

是否是钓鱼邮件：否

置信度：20%

- 相同输入，模型超参数固定（do_sample: false, temperature: 0, top_p: 1.0），DeepSeek-R1-Distill-Qwen-32B 输出存在一定随机性，即研判结论会跳变。
- 初步分析，主要原因是思维链不够稳定导致预测结果控制不住，DS后续版本可能会进行优化。

02

产品战略方向

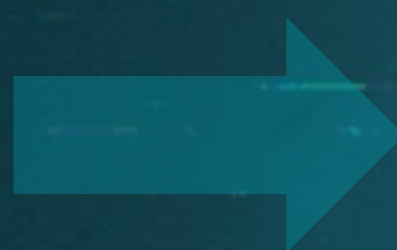


AI时代对安全供应商的要求

TVP Tencent Cloud
Valuable Professional

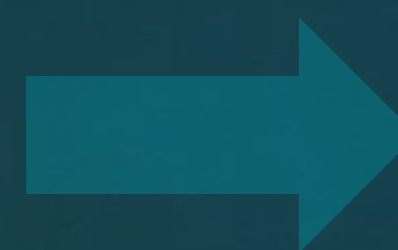
成熟产品

单品、封闭



解决方案

平台、体系

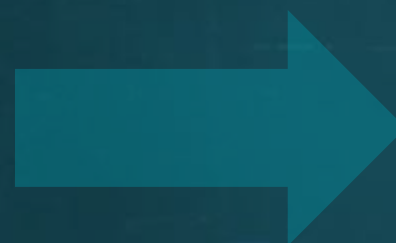


开放能力

语料、工程化

Work Flow

本质还是封闭任务



Agentic Flow/AI Agent

安全的manus

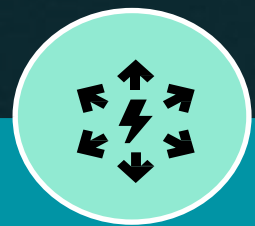


AI安全平台规划思路-整体技术框架





价值落地交付方式



① 使用

直接采购和使用现有安全大模型

示例：直接采购和使用安全运营大模型、钓鱼威胁检测大模型、数据安全大模型

使用现有模型服务



② 嵌入

通过API集成安全大模型到现有平台

示例：将钓鱼威胁检测大模型推理结果以API方式集成至现有安全运营平台

调用现有模型服务

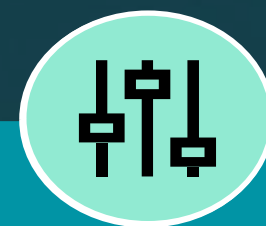


③ 扩展

通过RAG来扩展安全大模型专业领域知识

示例：将用户资产信息导入安全运营大模型，基于RAG让大模型有效利用用户资产信息进行准确的研判和输出

智能体自定义与开发



④ 定制

通过微调来扩展安全大模型专业领域知识

示例：基于用户工控物联网有限的有监督数据集和安全基础大模型，微调出用户工控物联网安全威胁检测大模型

大模型微调



⑤ 新建

从零开始构建自定义安全垂域大模型

示例：全流程实施数据工程、预训练、微调、训推优化、智能体开发工作，探索构建全新安全垂域大模型[不推荐]

全流程大模型训练

可控性和灵活性

能力要求、资源要求、总成本

03

核心技术介绍



深度融合产品的智能体快速构建技术

对安全智能体的理解：基于安全数据感知，自主决策能解决复杂场景问题，并且能闭环；
内置5个超级智能体场景，专家模式300多种安全插件灵活编排。

AI共创

我的AI

全览

快捷

数据管理

知识库

近义词库

XDR

专有名词库

ZTP

微调数据集

智能评估

DSP

评估任务

测试数据集

模型管理

模型仓库

模型训练

模型加密

模型压缩

我的应用

我的应用

新增

名称/描述

自动驾驶RAG增强智能体

已发布

基于生成检索增强RAG技术，通过用户企业内网实际资产详情增强安全GPT自动驾驶的思维链研判...

大模型提示注入检测智能体

已发布

专门用于识别和防范提示注入攻击。该智能体可检测应用输入/输出内容中的异常模式，保护用户数...

辅助驾驶文件问答RAG增强智能体

未发布

介绍信息介绍信息介绍信息介绍信息介绍信息介绍信息介绍信息介绍信息介绍信息介绍信息介绍信...

资产安全管理智能体

未发布

通过实时数据分析和风险评估，该智能体能够识别潜在威胁，提供安全建议，并确保资产在整个生...

生产网安全运营智能体

未发布

基于工业互联网实时监控和智能分析，该智能体能够识别网络入侵、异常流量和潜在威胁，确保生...

设备运行状态分析智能体

未发布

通过分析设备运行数据，该智能体能够预测潜在故障，提供维护建议，帮助企业优化设备管理，降...

邮件防泄密智能体

未发布

通过智能分析邮件内容和附件，该智能体能够识别敏感信息并防止未经授权的泄露，确保企业数据...

安全合规咨询智能体

未发布

基于工业互联网实时监控和智能分析，该智能体能够识别网络入侵、异常流量和潜在威胁，确保生...

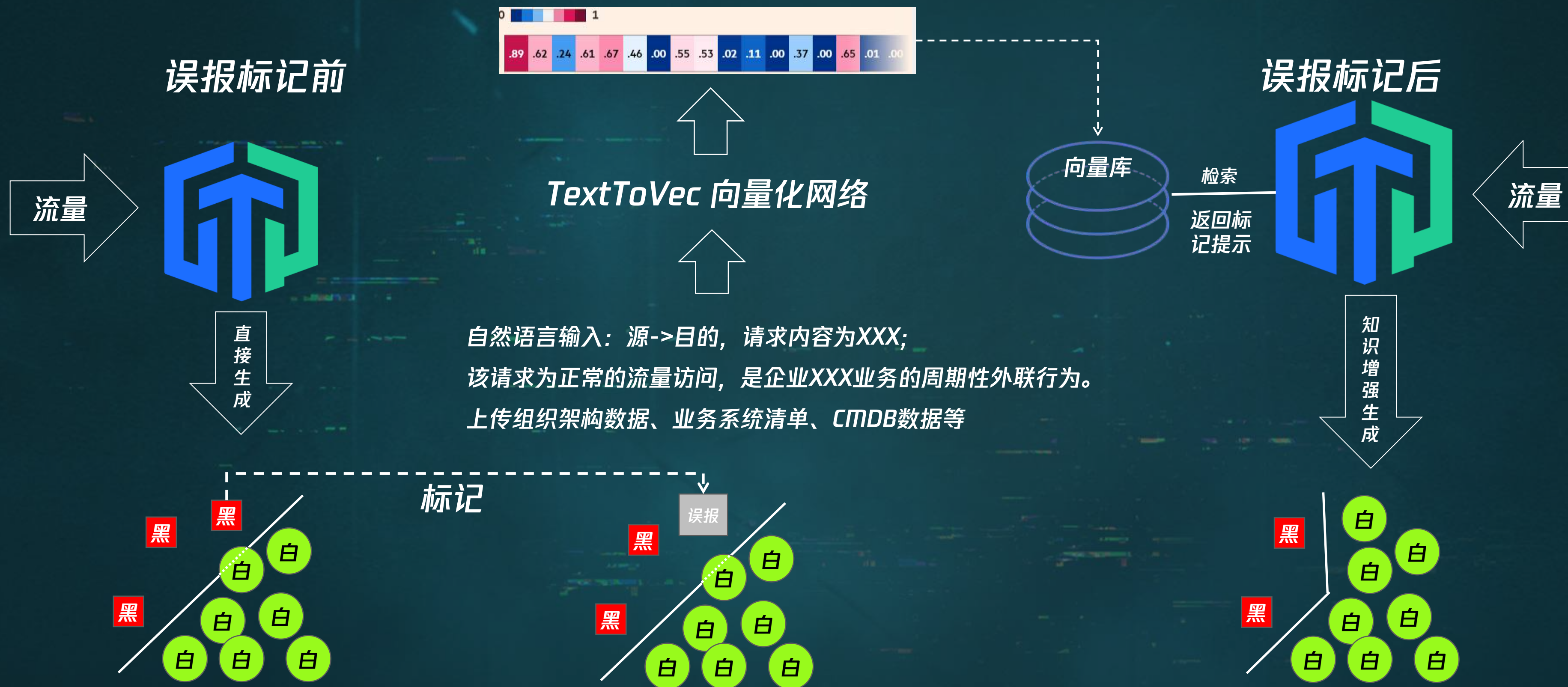
代码安全分析智能体

未发布

检测软件代码中的安全漏洞和潜在风险。该智能体能够提供详细的安全报告和修复建议，帮助开发...



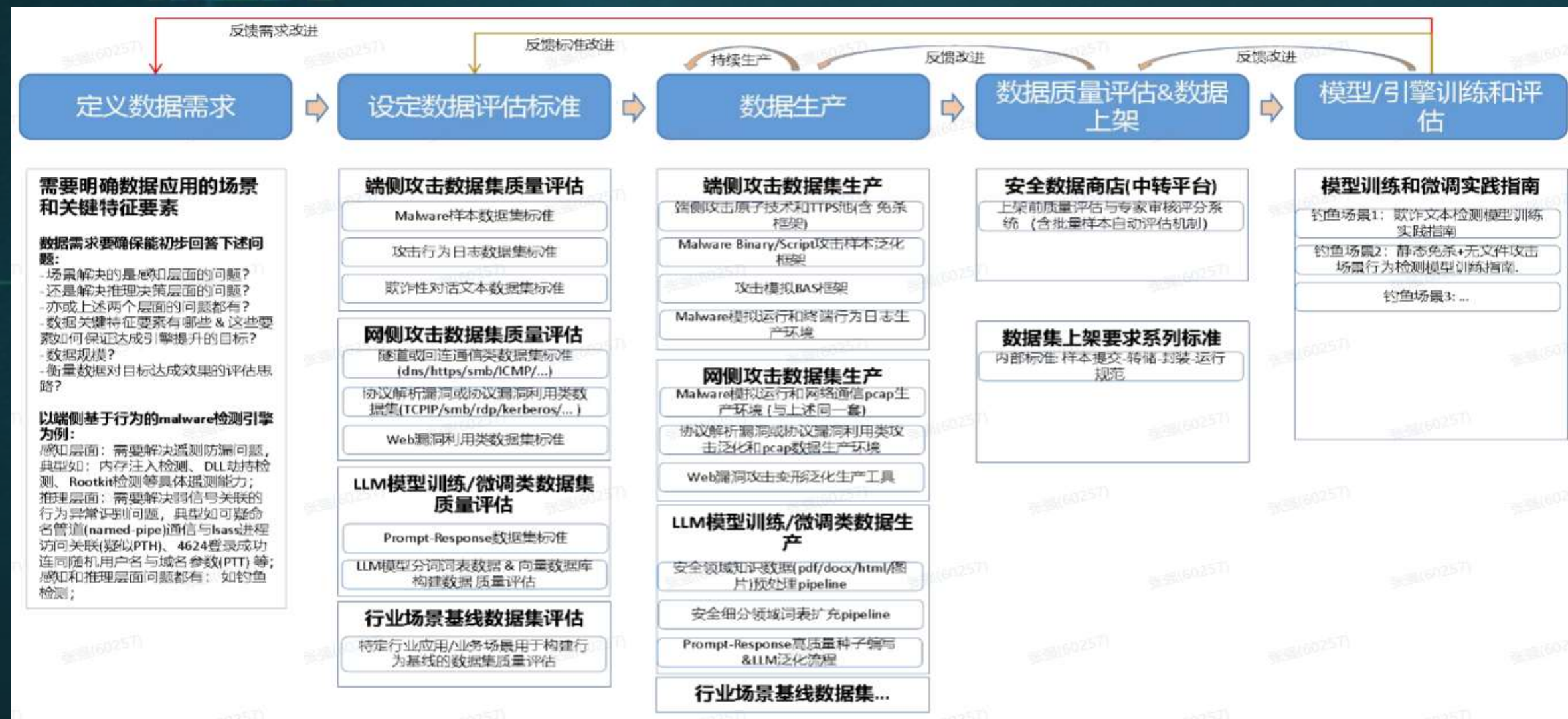
基于RAG技术的灵活调优





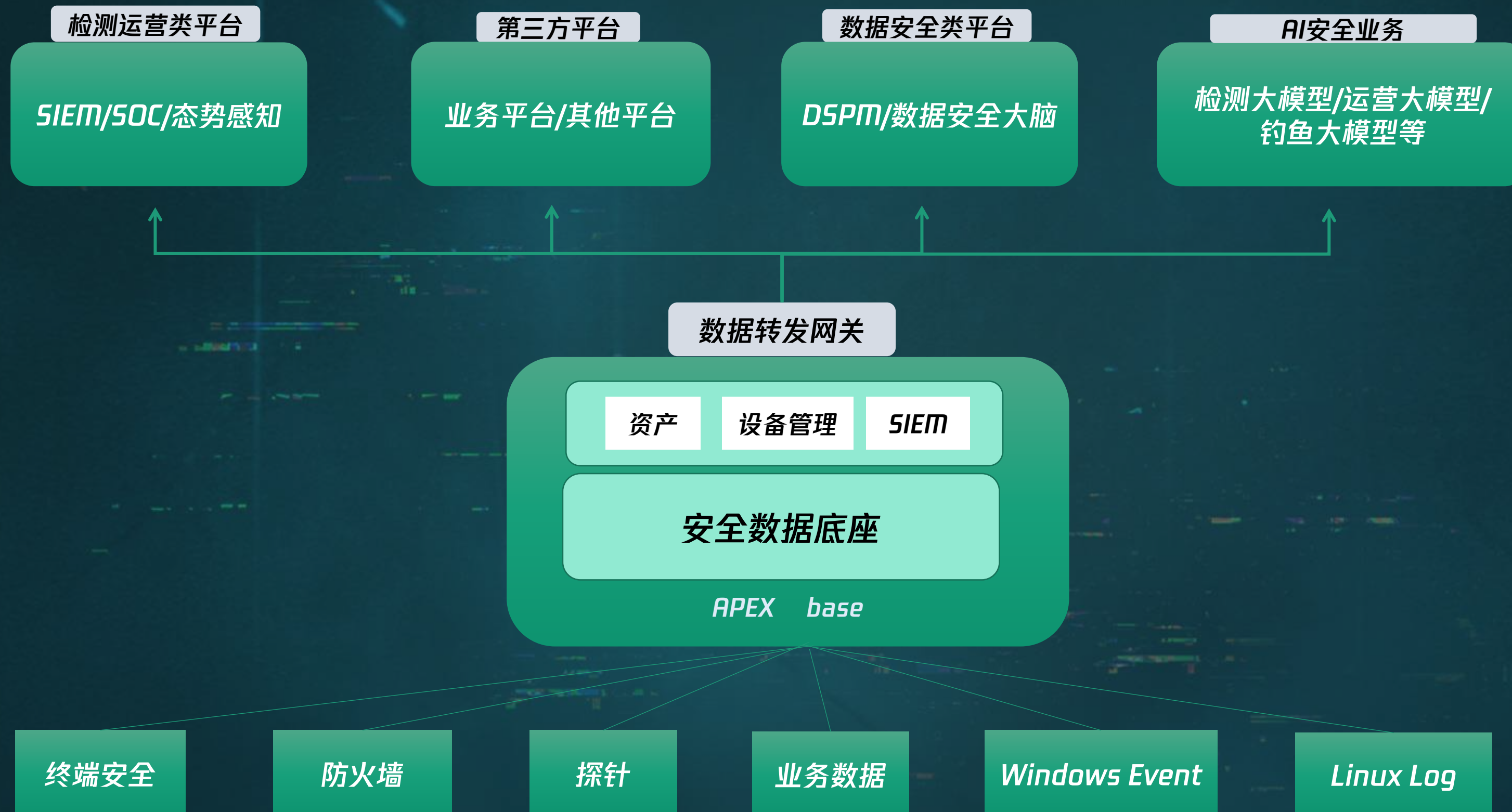
安全语料工程化技术

- 1.深信服蓝军将多年积累近400种有效的高对抗性实战攻击基础手法，通过脚本化改造，形成攻击Payload的种子样本。
- 2.借助DeepSeek对这些种子样本进行随机变形，泛化出更多的高对抗原子攻击样本。
- 3.通过自动化工程，对这些高对抗原子攻击样本进行多种形式的排列组合，与业务流量整合后形成海量的蕴含实战攻防对抗经验的攻击流量。
- 4.所有的攻击流量都会送入靶机系统进行自动化测试，对不合格的攻击样本〔比如不符合基本业务逻辑〕进行修正和强化。
- 5.合格的攻击样本进一步发送给安全检测引擎进行检测，保障和加强攻击样本强度。
- 6.通过“脚本构建-测试-筛选-修正”的自动化迭代过程，最终会形成大量实战对抗样本，这些数据将会应用于安全大模型的预训练、微调 and 测试。





网数融合、AI赋能的统一数据底座架构



大模型推训技术—最懂AI推训的安全厂商

TVP Tencent Cloud Valuable Professional

企业应用场景

文案生成助手

企业知识问答助手

商机分析助手

对话式应用改造

智能客服助手

经营分析助手

ISV应用承载

深信服安全GPT

卫宁WiNGPT

AI应用创新平台

数据模块

企业数据接入
长记忆模块

提示工程

prompt模板
prompt优化

企业知识库

文档向量化
检索能力增强

workflow编排

低代码编排
内置最佳实践模块

评估优化

应用效果评估
问题辅助分析



模型仓库



数据管理



训练自动化



推理优化



模型加密

AI基础设施

异构算力资源管理

容器引擎

高性能网络

分布式高性能存储

主流硬件



国际厂商GPU



国内厂商NPU



国内厂商GPU



CPU



Network

HCI资源池

大模型训推一体机

私有化部署

托管云

企业AI应用市场

RAG加速服务

04

关键场景实践



已经成熟商业化的实践场景

TVP Tencent Cloud
Valuable Professional

唯一经历过2024HW实战的AI安全厂商

1

威胁检测大模型

高对抗、高混淆Web攻击威胁与攻击成功检测，如 Web 0-Day利用、Webshell 文件上传与通信加密、SQL/Java/PHP各类注入攻击、XSS攻击、Java反序列化攻击等；
基于自然语言的Web攻击解读，包括攻击意图、Payload分析、攻击结果判断、处置建议等

3

钓鱼邮件大模型

高对抗、高混淆钓鱼邮件威胁与攻击检测；
基于自然语言的攻击解读，如邮件文本内容的欺骗和诱导意图、二维码和外链的网页代码分析、加密附件内容，还原攻击过程

6

行为分析大模型

业务访问权限设置的推理研判，实现权限可视以及常态化、持续性权限风险监测；
基于自然语言的访问权限和访问行为风险解读，如终端失陷后横向异常访问、员工跨业务异常访问、权限配置与岗位不符等

2

安全运营大模型

全量告警自动研判，聚合、去冗、降噪、分类定性；
基于自然语言的运营助手，如态势解读、情报分析、威胁研判、资产漏洞、安全百科等。SOC+GPT=AI SOC

5

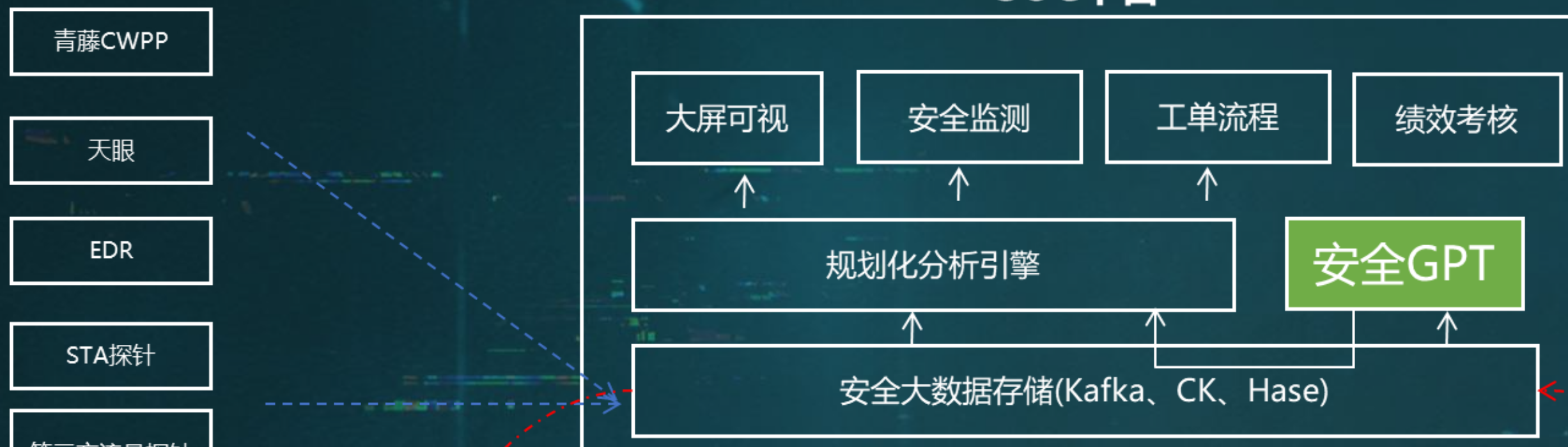
数据安全大模型

实时访问中的数据类别级别梳理和研判，自动识别其中的敏感类数据；
基于自然语言的数据资产分布、敏感数据访问、意图和风险推理等解读，如短期内大量数据访问、数据API权限不合理等



现有SOC+安全GPT=AI SOC

SOC平台

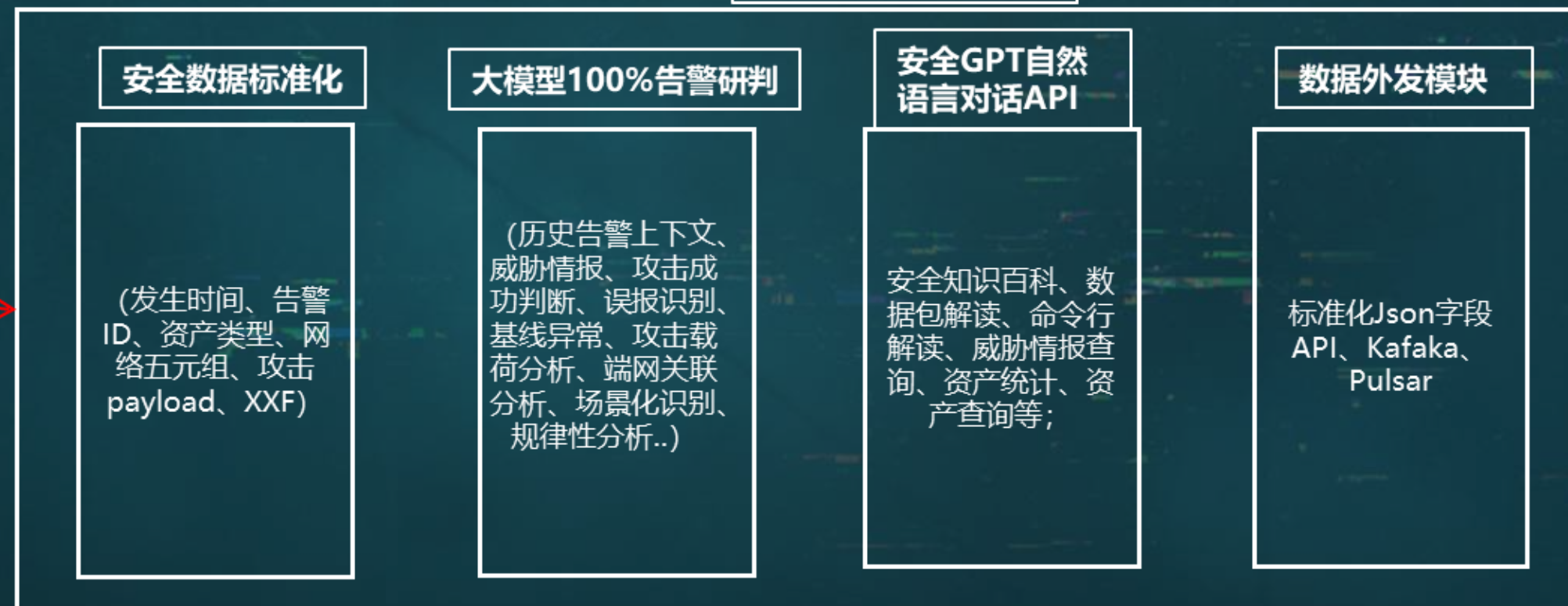


100+安全告警(带研判结果、定性标注)

业务误报、有效攻击、病毒木马、脆弱性风险、外部攻击未成功

安全GPT

10000+原始安全告警





不止做产品供应商，联合共建AI人才培养体系

《AI安全技术与应用》课程大纲

1. AI大模型基础

- 生成式AI
- 大模型应用项目生命周期

2. 大模型安全防护

- 大模型主要安全风险
- 大模型攻击典型手法
- 大模型安全防护手段

3. 安全AI应用实践

- AI应用到安全的发展概述
- AI落地安全运营的理念与实战
- AI落地数据安全的理念与实战
- AI落地流量检测的理念与实战
- AI落地钓鱼邮件检测的理念与实战
- AI自定义应用构建——模型微调
- AI自定义应用构建——智能体构建

AI网络安全人才培养

AI安全治理规范的修订与推广

AI安全创新文化沙龙建设

国标、行标与创新课题联合申报

业务场景

- 安全部门看到漏洞/入侵，需要和业务部门排查沟通，跟进工单，耗费大量精力占用〔安全人员三分之一的精力〕
- 钓鱼邮件演练效果很好，但难以覆盖到每个人，如中招人员比较多，涉及的沟通确认和安全意识培训的工作量巨大
- UEBA检测到人的异常行为，关键在于出现异常后，必须和出现异常的终端用户沟通确认，工作量巨大

落地价值

- 极大释放安全人员精力，代替安全人员与普通员工、IT人员进行日常沟通确认，是真正把安全智能化往前推一大步的核心环节。
- 直接提升事件处置、告警研判、安全意识替身、ueba检测等安全工作的闭环效果，帮助安全人员把工作做得更好。
- 凸显安全团队价值，让普通员工高频感知到安全团队的工作，普通员工也可进行管理制度、安全常识问答。

发现高危漏洞



自动调用漏扫API检测全局漏洞风险



助手自动通知责任人整改



整改后生成复核工具



自动生成整改报告



智能体场景实践：HW值守智能体

业务场景



落地价值

8点半自动完成
巡检报告

事件分析，自动导入
查询语法

自动告警分析

自动封堵提报

自动生成
值守报告



到岗即查看巡检报告，
大幅节约琐碎的巡检时间



语法自动通过智能体查询，
异常告警输出报告



封堵自动提报



面向未来：单体安全应用向安全AI原生应用演进



计算资源

单体安全应用

支撑安全运维、安全运营等业务的单体安全应用，用于统一运维管理安全组件，开展识别、监测与响应等工作，可能会通过接入大模型提升应用运行效率、优化运营效果。

传统面向人类的工具

传统工具通常不合适直接面向机器人直接使用



计算与智算融合底座

安全AI原生应用

以AI为出发点，将智能体作为核心驱动力，重新设计和构建应用，逐步以智能体的管理、监督为主要用户界面，逐渐替代传统UI，推动安全业务稳步实现全程化托管。

人类监督机器人开展工作

面向机器人的工具改造+为人类监督所设计的UI+机器人



智算资源

安全智能体

具备感知安全环境，数据存储和一定的行为能力，甚至有一定的自主性，能分析决策并执行任务，安全部门可以基于业务需求，编排、优化、构建的多元化AI安全智能体。

以大模型为大脑的机器人

在有大脑思维能力的基础上，还需要集合新知识与行为能力

THANKS

谢谢观看