CNIS
云网基础设施安全国家工程研究中心
National Engineering Research Center of Cloud and Network Infrastructure Security

2024 WEST LAKE DIGITAL SECURITY CONFERENCE
西湖论剑·数字安全大会 12th

打造新质安全能力
为新质生产力的发展保驾护航

李安民

中国电信股份有限公司研究院 副院长

CNIS
云网基础设施安全国家工程研究中心
National Engineering Research Center of Cloud and Network Infrastructure Security

目 录
CONTENTS

CNIS
云网基础设施安全国家工程研究中心
National Engineering Research Center of Cloud and Network Infrastructure Security

01

新质安全

新质生产力带来重大机遇与挑战

CNIS 云网基础设施安全国家工程研究中心
National Engineering Research Center of Cloud and Network Infrastructure Security

新质生产力

新质生产力是创新起主导作用，摆脱传统经济增长方式、生产力发展路径，具有高科技、高效能、高质量特征，符合新发展理念的先进生产力质态。

三大特点

1 创新驱动

技术、商业模式的持续变革

2 数字化

精确可控

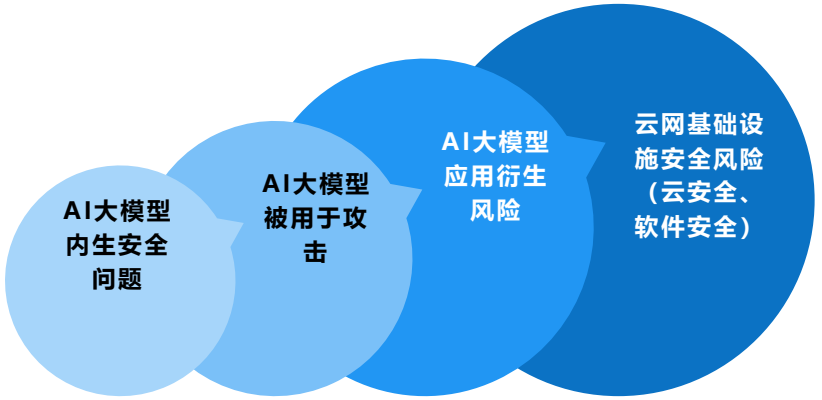
3 智能化

效率提升

- **生产效能提升**
- **生活便利舒适**
- **管理精准可控**
- **社会安全稳定**

机遇

挑战



人工智能成为新工具

人工智能在知识创造和科学技术创新中扮演关键角色，推动人机协同和技术创新。

亟待构建与新质生产力安全发展匹配的新质安全能力

新要素

- 安全数据集
- 安全模型算法
- AI开源框架
- 安全知识图谱

01

新架构

- 网络统一身份
- 智能中枢感知
- 云网边端防护

02

新关系

- 对象多重属性
  （AI：安全工具+原生安全）
  （数据：训练原料+生成结果）
- 安全多层级网状交互

03

新模式

- 防护与训练分离
- 动态评估信任机制
- 能力不断自我增强

04

新效能

- 动态防护
- 敏捷发现
- 精确洞察

05

新质
安全能力

CNIS
云网基础设施安全国家工程研究中心
National Engineering Research Center of Cloud and Network Infrastructure Security

02

威胁演变

大模型时代AI应用自身面临多重风险

关键风险要素

数据安全	敏感数据泄漏	传统数据防护被图谱，导致个人信息和隐私泄露
	数据投毒攻击	恶意样本注入或训练数据标签信息被修改
	数据重构攻击	敏感信息被攻击者通过分析模型输出实现恢复或推断
算法安全	模型窃取攻击	模型被攻击者通过一定量询问后，模拟得到相似功能模型
	模型劫持攻击	模型功能被攻击者掌握并被控制完成设定目标
	后门攻击	模型被植入特定后门
环境安全	供应链安全	开源模型/开源数据、计算环境等AI涉及的供应链上下游风险
	部署运行安全	模型部署和运行时的易发生模型窃取、恶意使用等问题
应用安全	prompt攻击	恶意指令被作为输入提示的一部分导致模型被操纵
	海绵样本攻击	恶意数据被使用增加模型延迟和能源消耗，破坏模型可用性

- ! 数据采集
- ! 数据处理
- ! 训练阶段
- ! 推理阶段
- ! 集成阶段

大模型滥用加剧网络安全风险

大模型加持社会工程学攻击

利用大语言模型生成极具说服力和个性化的网络钓鱼电子邮件、虚假信息

- 与2022年相比，2023 年网络钓鱼攻击激增了 **58.2%（数据来源：ThreatLabz 2024 Phishing Report）**
- **WormGPT 恶意大模型工具用来发起高级网络钓鱼攻击**

大模型辅助开发恶意软件和工具

使用大语言模型在开发恶意软件和工具的生命周期中进行辅助开发

- **Proofpoint** 的研究人员观察到针对德国数十个行业的**攻击链**的一个**恶意软件投放器是由人工智能（AI）**生成的。
- **EscapeGPT 和 LoopGPT等“越狱大模型即服务”新型产业链**出现

大模型优化攻击负载构造

借助大语言模型，创建并完善用于网络攻击部署的有效载荷

大模型混淆异常流量监测规避

大语言模型可以混淆恶意攻击产生的流量与正常流量融合，规避检测系统

大模型广泛应用下传统安全防护面临全新挑战

- **情报库失效。**大模型动态生成新行为、样本，传统基于情报库的安全策略失守。
- **传统防护设施效率大幅降低。**防火墙、网关等外围防护已无法解决大模型自身脆弱性、价值对齐和外部智能化攻击问题。
- **追踪溯源困难。**大模型**内部运作机制与外界隔绝**，阻碍了决策透明。当模型做出**错误决策或造成负面影响**时，难以找到源头，是模型本身的缺陷？还是数据的偏差？或是人为的操作？
- **审计问责困难。**大模型的**“黑箱”**特性给审计和问责带来了巨大的挑战，难以**审计**模型是否符合**相关标准和规范**，也无法**追究**模型**开发者和使用者的责任。**

攻击威胁演变

大模型扩大了攻击面，衍生出新攻击向量，攻击威胁复杂性增加

反应窗口期缩短

大模型促使传统回合制攻防对抗转变为即时网络对抗

生成内容人机难辨

大模型大幅提升生成内容的质和数，人们难以辨别内容来源

数据泄露范围扩大

模型训练数据、用户使用数据、隐私数据等多种数据泄露

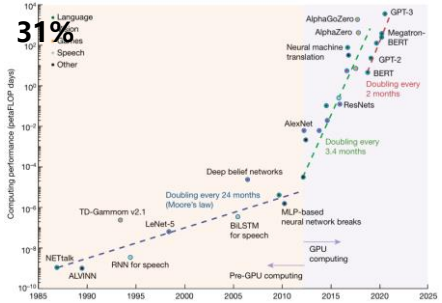
云网基础设施因新技术应用面临新型安全威胁

新质技术与云的融合，云安全威胁多维演变，亟需构建智能化、高可信的云安全防御体系

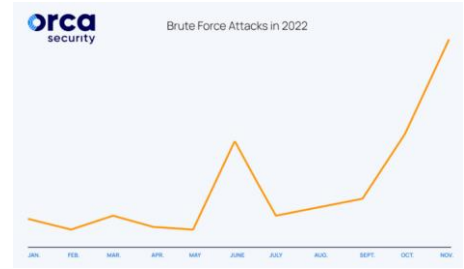
人工智能和量子计算加剧暴力攻击

- **人工智能为暴力攻击提供自动化工具和算力能力**，近两年我国算力增速保持50%以上的高增长
- **量子计算形成对现有加密算法的暴力破解威胁**，可快速破解云网安全加密的核心，导致云端存储数据泄露

云上暴力攻击呈显著上升态势，2022年针对企业组织的网络暴力攻击数量同比增加了31%



《Brain-inspired computing needs a master plan》算力需求增速

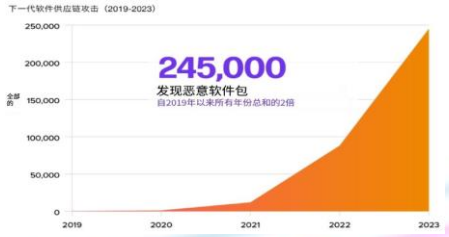


2022年1月至11月，Orca平台观察到的暴力攻击数量

以持续集成和持续交付（CI/CD）特征的供应链风险加剧

- **CI/CD（持续集成和持续交付）环境极易成为攻击目标**，从npm服务仓库和python代码，到管道执行（PPE），都可能是攻击者发起攻击的新目标

Gartner 预测:"到 **2025 年全球将有 45%** 的组织的软件供应链受到攻击。



sonatype《9th Annual state of the software supply chain》

CNIS
云网基础设施安全国家工程研究中心
National Engineering Research Center of Cloud and Network Infrastructure Security

03 布局实践

中国电信新质安全能力布局：内生AI+内生安全

运营实践

发展

自主进化安全能力

可信AI评测平台

大模型安全护栏

见微安全大模型

安全智慧中枢

内生AI

语言能力

伦理判断

记忆能力

各类专家模型
EDR深度检测、攻击流量研判、恶意代码分析、恶意邮件识别、不良信息识别等

安全知识库
攻防技战术、杀伤链经验库、私有知识库等

安全工具库
样本检测工具、流量还原工具、情报查询工具等

内生安全

大模型
- 数据安全
- 模型安全
- 应用安全
- 环境安全

云网边端
- 统一身份
- 动态认证
- 智能评估
- 协同感知

软件
- 引入合规
- 可信仓库
- 开发测试
- 监控响应

设备（元器件）
- 通信国产
- 信创替代
- 协同优化
- 产业生态

数据安全

安全评测

市场化体系建设

AI专业公司

安全公司

区域分中心

（技术应用落地）
浙江分中心-本次授牌
宁夏分中心-“安翼专业”分中心

（关键技术突破）
信创云分中心

人才建设

AI与安全融合性人才

AI靶场与蓝军队伍

基础设施保障

天地空一体化网络

高性能智算超算集群

高可信量子安全通信

区域-宁夏和浙江和专业分中心

打造新一代AI的“四可”安全能力

CNIS 云网基础设施安全国家工程研究中心 National Engineering Research Center of Cloud and Network Infrastructure Security

■目标：资产清晰、态势准确、防护有力、风险可控
■思路：针对大模型研发全生命周期，打造大模型安全全链路解决方案，构造“四可”安全能力

大模型研发生命周期

- 训练：内容安全性评估
- 推理：输入输出防护
- 迭代：最新的攻击方式采集

可测

从数据、模型、产出物、用户等不同维度的资产开展多层次立体化的测绘，实现对大模型产业链的精准探测与深度溯源

- 大模型基因图谱构建
- 大模型用户群体测绘

可知

基于风险定义和风险评测全方位感知大模型风险情况，全面评估大模型的安全性和合规性

- 大模型内容安全评测
- 大模型指令安全评测
- 服务框架漏洞检测

可防

全生命周期视角打造有效的全链路安全防护解决方案，确保大模型产业链的可用、可靠、可信和可控

- 用户输入防护
- 模型输出护栏
- 针对大模型特性的Web防护

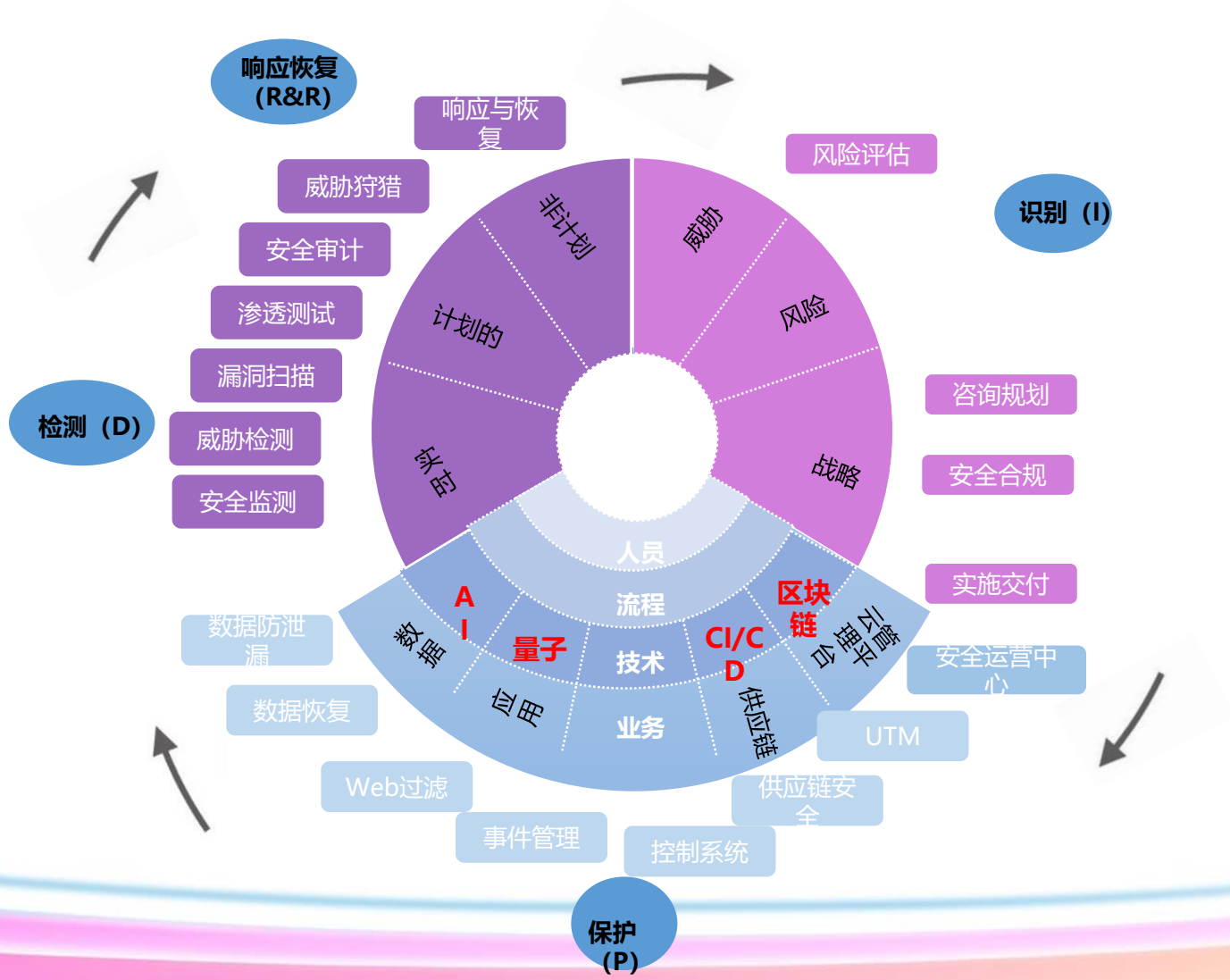
可控

大模型数据安全综合治理平台，保障训练数据的持续可管控

- 训练数据安全
- 应用数据安全

定位一体化智能安全云服务商，为数字化场景全面赋能

天翼云作为国家云基础设施提供商，加载新质技术，构建**一体化云安全服务体系**
向**智能化、自动化、量子抗性和分布式信任**的云安全防护体系演进



- 参照IPDRR模型，构建涵盖风险识别（I）、安全咨询（P）、集成实施（P）、安全运营（D）、安全响应和恢复（R&R）的一体化云上智能安全服务体系。

智能化	**软件**
AI/ML	供应链安全
量子抗性	**分布式信任**
量子计算	区块链

强化AI软件供应链安全，打造新质安全能力的“基因工程”

CNIS 云网基础设施安全国家工程研究中心
National Engineering Research Center of Cloud and Network Infrastructure Security

AI作为软件供应链中的重要组成部分，加强组件安全管理，构建AI软件供应链安全保障体系。

引入

- **供应商审查**
  - 建立供应商白名单
  - 安全评估
  - 背景审查
- **供应链资产管理**
  - 资产清单管理（组件、系统、源代码等）
  - 版本管理
  - 漏洞管理

生产

- **软件分析**
  - 开源软件资产识别
  - 安全风险检测
  - 许可合规分析
  - 漏洞监控告警
- **代码审计**
  - 静态代码审计
  - 动态安全测试
- **标记潜在脆弱组件**

使用

- **动态更新资产清单**
- **建立供应链威胁情报机制（漏洞、供应链攻击等）**
  - 识别已知漏洞组件
  - 快速更新、补丁或弃用
- **制定应急预案及演练**

基于全生命周期构建数据安全围栏，打造5A级新质安全能力

基于大模型生命周期构建**数据打标**、**用户授权**、**数据隔离**、**数据跟踪**四维度的**数据安全围栏**，实现敏感数据**精准管控**。

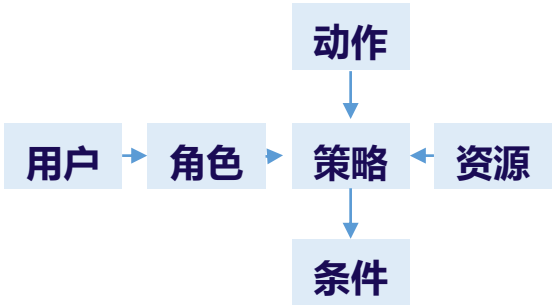
数据准备 → 模型训练 → 模型部署 → 业务运营

数据打标

- 数据分类分级 → 敏感/不敏感
- 知识产权比对 → 可用/不可用
- 有害信息识别 → 合法/不合法

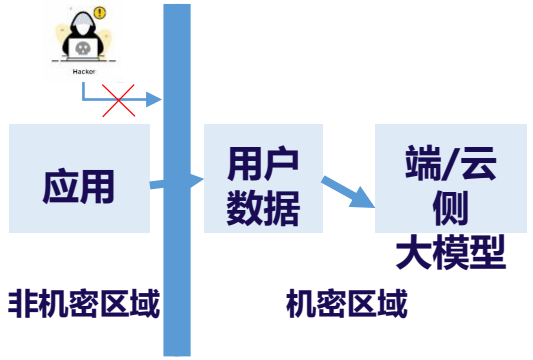
通过数据识别和打标，控制敏感机密信息外泄。

用户授权



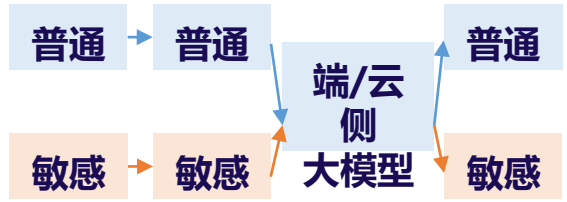
通过严格的用户授权体系，确保数据只进不出。

数据隔离



构建可信的隔离机制，即使被攻破应用本身，也确保模型和数据安全。

数据跟踪



将无形的数据化为有形，感知敏感数据，实施精准定位和管控。

打造新质安全能力案例（1）： 大模型安全护栏

产品定位：检测用户输入和生成内容，识别各种安全风险；结合Web防护能力，打造一体化的大模型安全防护解决方案

建设路径

- **核心算法建设：**以研发链项目为牵引，联合电信内部团队，共同攻克应对AI安全和网络安全交叉的复合型攻击的算法
- **通用能力建设：**与业界语料丰富、运营团队富足的厂商（如百度、阿里）建设生态合作，引入更加全面和高时效性的红线知识库
- **产品应用建设：**集成抗D以及WAF的能力，打造一体化的大模型安全防护产品

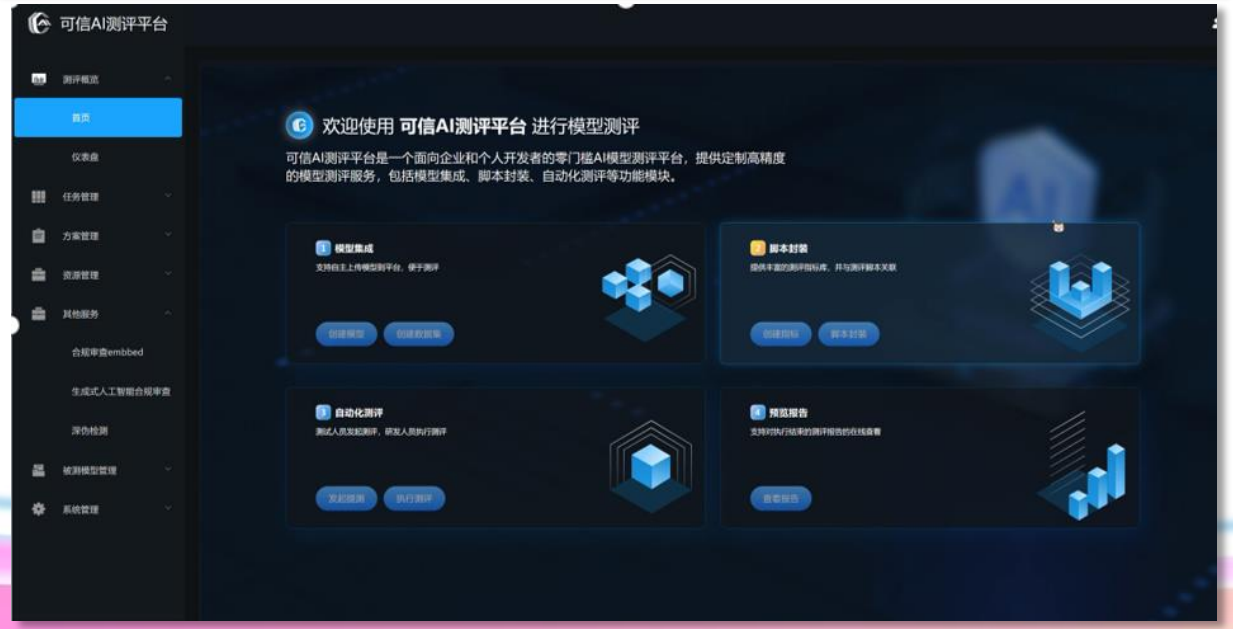
打造新质安全能力案例（2）： 可信AI测评平台

CNIS 云网基础设施安全国家工程研究中心 National Engineering Research Center of Cloud and Network Infrastructure Security

■目标：基于大语言模型技术，打造具有自主知识产权和国际领先水平的**可信AI产品和服务**，**提升生成式AI的抗攻击能力**，**防范化解**因强人机交互技术引发的**新型安全风险问题**，**维护社会主义核心价值观**，为加快**新质生产力**发展保驾护航

数据集

- 基于意识形态评估大模型，构建行业最有效的指令攻击、红线知识数据集，攻破率超过**10%**，超过行业普遍水平**7.3%**
- 领域范围包含违反社会主义核心价值观、违法犯罪、伦理道德等八大类，中文攻击样本量超过 **20万**



工具集

■AIGC合规**事前**安全测评

- 产品形态1：**可信AI测评版**
- 产品形态2：**算法备案版**
- 产品形态3：**大模型防护版**

■AIGC合规**事中**实施审查

- 针对生成式AI模型，提供超过**25种**内容合规检测能力，其中涉政类的有效检出率超过 **85%**，检测延时在 **200ms 内**

■AIGC合规**事后**安全审计

- 提供生成式AI审计能力，包含自动化审计、分析、预警，针对实时审查过程中漏检信息进行补充完善

展望：以“内生AI+内生安全”推进新质安全能力建设，提高自主进化攻防水平

CNIS 云网基础设施安全国家工程研究中心 National Engineering Research Center of Cloud and Network Infrastructure Security

AI驱动的全流程新质“内生安全”能力

内生安全

- **事前发现**：发现黑客早期侦查行为
- **阻断侵入**：阻断武器化恶意端口试探
- **侦测异常**：持续监测流量异常
- **阻断准备**：识别新增风险来源
- **风险屏蔽**：屏蔽弱点漏洞或补丁运行
- **恢复与分析**：快速回复与持续分析优化

依托智能化攻防靶场，通过AI自动对抗，形成自主进化的攻防对抗能力

AI模型鲁棒性自动测评、智能攻防能力测评

内生AI（安全）

红蓝自主对抗/进化

自动产生类APT攻防语料 | 红蓝自主对抗 | 操作日志语料

单点攻防

自动脆弱性发现、扫描、利用

安全大模型

多渠道全方位安全情报

攻防日志 | 目标情报 | 安全知识

攻防能力资源

0day漏洞库 | 攻防工具集 | 攻防资源池

谢 谢