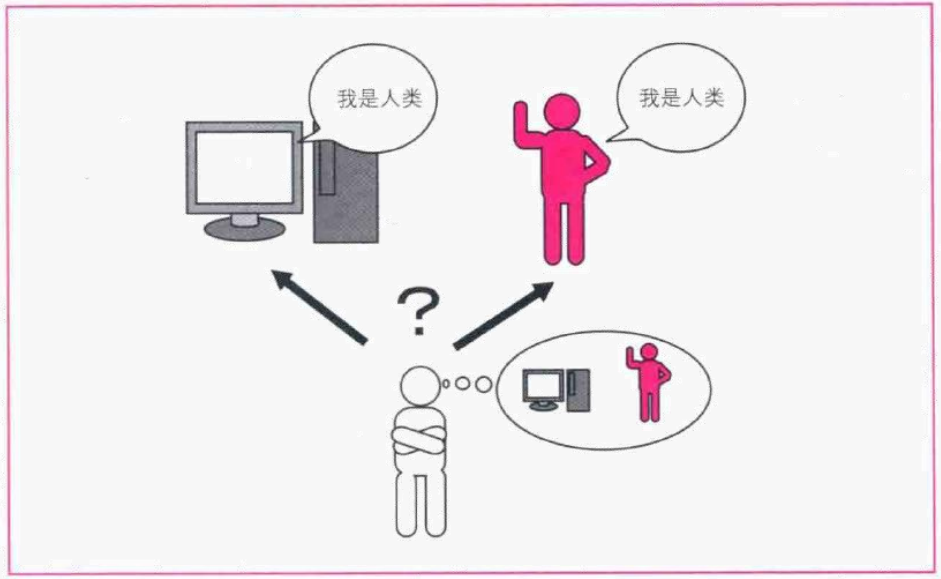
证券代码：002212

人工智能与安全的融合

天融信科技集团

人工智能的诞生：图灵测试



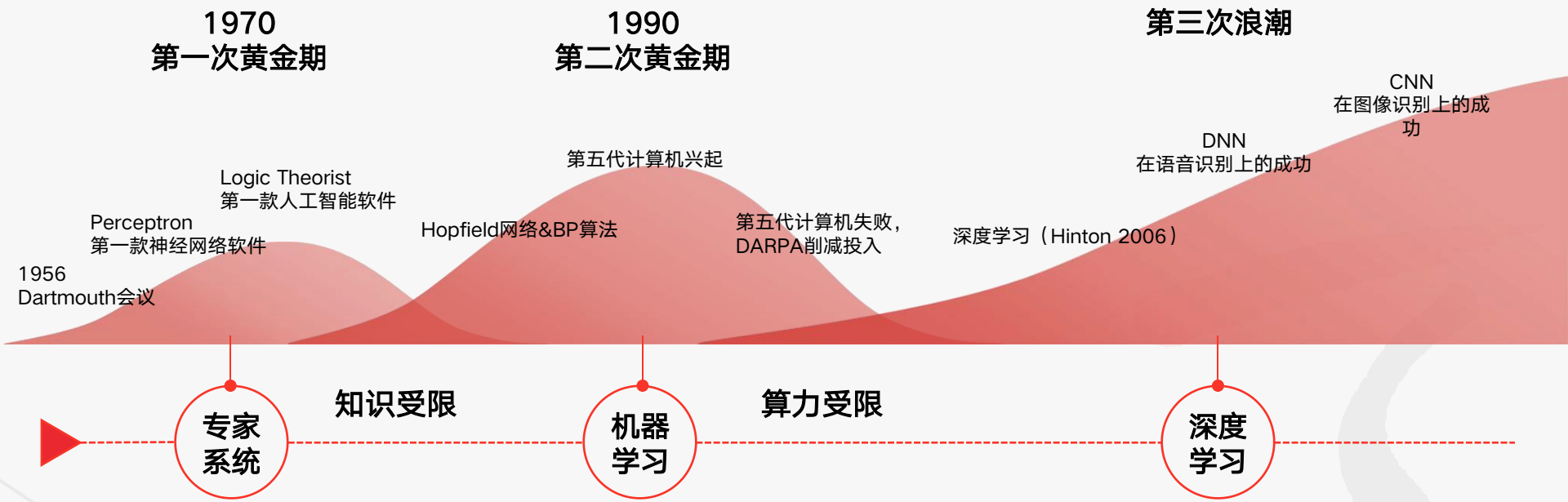
图灵在1950年发表了论文《计算机器与智能》。在这篇论文中，他对人工智能的发展与人工智能的哲学进行了深刻的讨论。

图灵测试：

将测试者与被测试者隔离，为了避免机器的声音影响测试结果，测词者只通过键盘和显示器等设备以文字形式向被测试者提问，然后判断对是人还是机器。在2014年的图灵测试大会上，一台俄罗斯的超级计算机伪装成13岁的男孩，回答了测试者输人的所有问题。其中有33%的测试者认为与自己对话的是人而非机器

人工智能 = 人为地使设备或软件模仿人类的行为

人工智能的三次浪潮



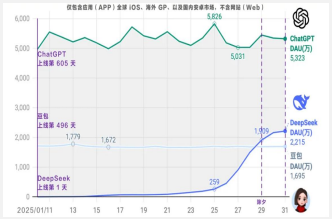
随着核心技术、数据量、场景和社会认知的就绪，A.I.产业迎来爆发的前夜！

DeepSeek的出现，加速了企业AI创新的动力

- DeepSeek 横空出世，强势改写人工智能领域的整体格局。凭借多维度核心优势，在业界掀起创新热潮，带来广泛积极影响，为国家科技发展注入强劲动力，同时也给各行业带来了机遇

急剧增长的用户规模

国内大模型全球用户占比最高



日活用户规模破3000

超160余家企业完成接入工作和产品化封装服务

技术竞争增强

实现技术指标反超，打破西方模型的优势，并创新技术应用，为**技术发展提供新路径**

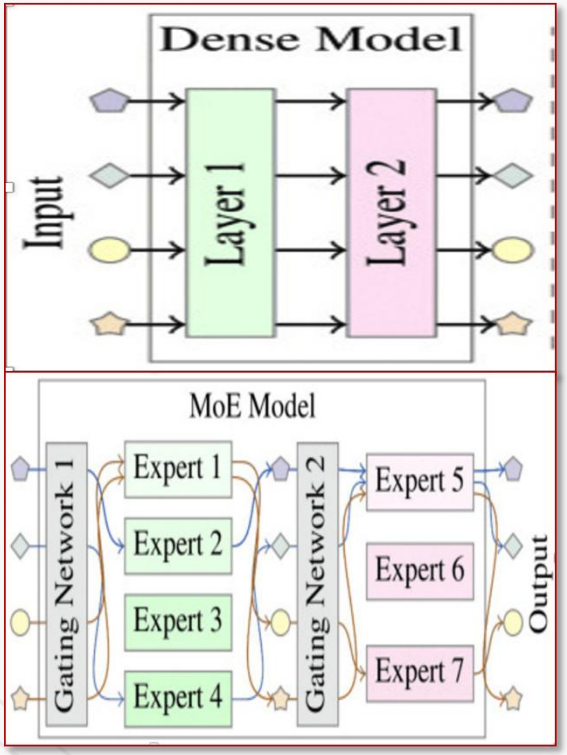
相对低廉的成本

通过“好用、开源、免费”的能力，降低了投入成本，提升了**行业的透明度和创新性**

行业认知格局开放

打破高投入、高资源才能投入AI领域的概念，凭借高效资源整合，激励更多用户投入**AI创新**

DeepSeek优势在哪里：Dense模型 vs MoE类型



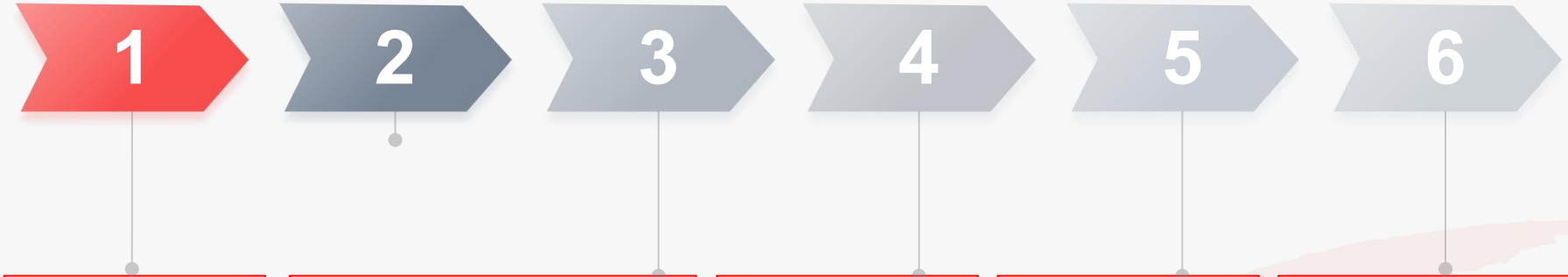
	Dense模型	MoE模型
结构	结构简单，每个神经元都参与计算	结构复杂，动态决定哪些神经元参与计算
训练	训练稳定，收敛性好，资源消耗大	动态选择机制复杂，可能不收敛，资源消耗小
推理	所有参数参与计算，资源消耗高	仅激活部分专家，计算高效
扩展性	参数线性增长，扩展成本高	通过增加专家横向扩展，成本可控
适应场景	简单任务、资源受限、强调稳定性	复杂任务、高效计算、极致性能
代表模型	Qwen2.5-72B、GLM、GTP-1/2/3、LLaMa	Deepseek-V3、DeepSeek-R1、GPT-4、Gemini1.5

DeepSeek-R1 vs 其他LLM

1. **提出**了一种新的强化学习算法
   GRPO
2. **尝试**了一种新的训练方法
   只用RL，不用SFT
3. **使用**了一种新的编程技术
   PTX

算力 = 硬件 + 软件

这么好的模型我们应用落地有哪些问题？



痛点一：方案立项阶段，结合业务需求，**用户难以预估需要使用的模型大小和训练、推理预计使用的资源量**，无法制作项目预算

痛点二：数据处理阶段，针对大模型开发的数据处理流程用户**缺乏最佳实践和自动化工具**包括需要准备什么数据集、按照何种格式准备场景任务的数据集、如何提升数据质量、在训练前按照什么样的配比抽取不同场景的数据等直接影响模型微调的效果，无法达到业务目标

痛点三：横型微调阶段，缺乏大模型开发经验，难以打通大模型微调的流程并优化过程中的资源消耗，导致大模型**微调训练难、训练效率低**

痛点四：模型评测阶段，业界缺乏统一的评测标准，不仅导致模型微调后不知道效果好不好，而且不确定模型的安全性是否有保障，**影响模型的开发效率**

痛点五：1）内容安全：大模型的输出内容相对不可控，如果不对输出的内容做出控制，**有可能存在监管、合规的风险模型安全**；2）模型安全：大模型开发成本高，若模型被窃取、算改，将对企业造成较大的**经济损失和影响**

天融信安全智算一体机

天融信安全智算一体机是以“算力硬件平台+智算平台”为基座，融合“计算、存储、网络、安全、智能”五大能力，内置DeepSeek R1模型，旨在为客户提供轻量化、高性能、安全可靠的一体化智算中心建设方案。

天融信安全智算一体机

	AI大模型支撑		核心业务承载		业务安全+模型安全	
安全智算平台	算力调度	模型推理	云主机	块存储	云主机安全	天问大模型
	模型仓库	模型训练/微调	容器	文件存储	网络安全	密码安全
	数据集仓库	模型评测	裸金属	对象存储	数据安全	运维安全
安全智算基础设施软件	异构算力资源管理		容器引擎		分布式高性能存储	高性能网络	内生安全网元
安全智算基础设施硬件	国际厂商GPU		国际厂商CPU		国内厂商CPU	国内厂商NPU	数通设备

主要功能

- **算力管理：**实现CPU、GPU资源的精细化分配和调度；
- **镜像仓库：**一键创建模型、虚拟机、容器；
- **模型服务：**内置多种主流大模型，满足各类场景需求；
- **AI模型安全：**全生命周期中有效保护数据隐私。
- **智算任务：**训练任务、推理任务、评测任务、量化任务

一体化算力底座

天融信天问智算云平台

AI 开发

模型训练 | 模型推理 | 模型评测 | 模型微调

运营+运维

租户管理 | 计量计费 | 运维监控 | 安全管控

算力调度

算力感知 | 独享算力 | 共享算力 | 算力优化

算力资源

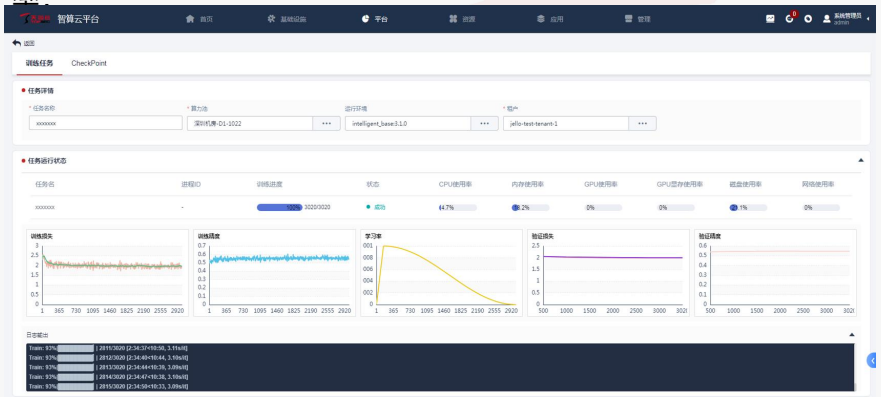
通用池

智算池：训练资源 | 推理资源

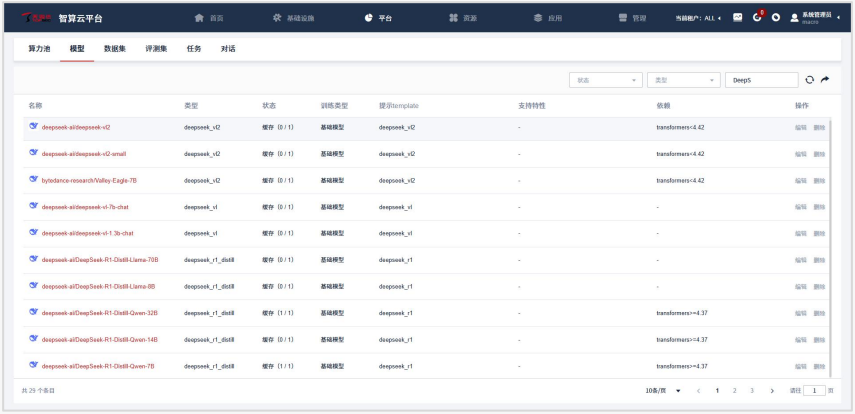
高性能存储 + 网络

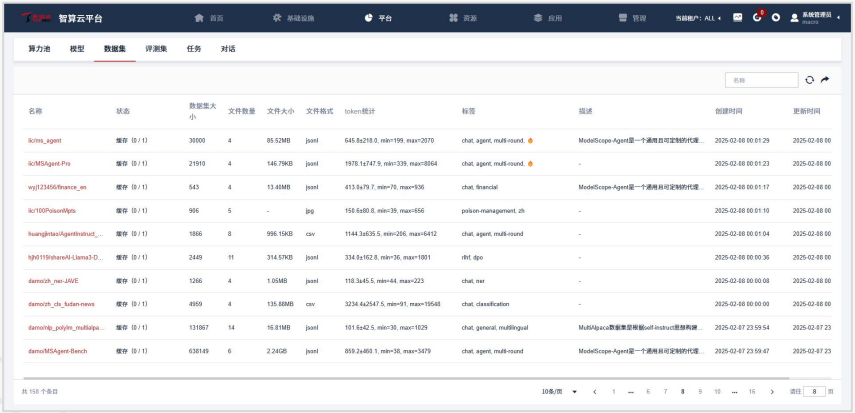
算力中心统一运营，提高管理效率

- 支持集群监控统计，健康巡检等的运维服务；
- 从平台架构、网络、服务层面保障租户的数据、算法、应用的资产安全；
- 从模型训练的全流程，提供可视化监控能力
- CPU/DPU/GPU 多芯片协同加速，自研高性能存储后端保障，支持算力亲和性调度，提升计算效率；
- 训练推理资源按需分配，动态调整，实现资源复用，提升资源利用率；



无缝接入DeepSeek生态，降低AI的应用门槛





支持DeepSeek全系列模型

- 覆盖从复杂决策到边缘推理的全场景需求

预置DeepSeek及其他模型相关的600+模型配置和150+数据集配置

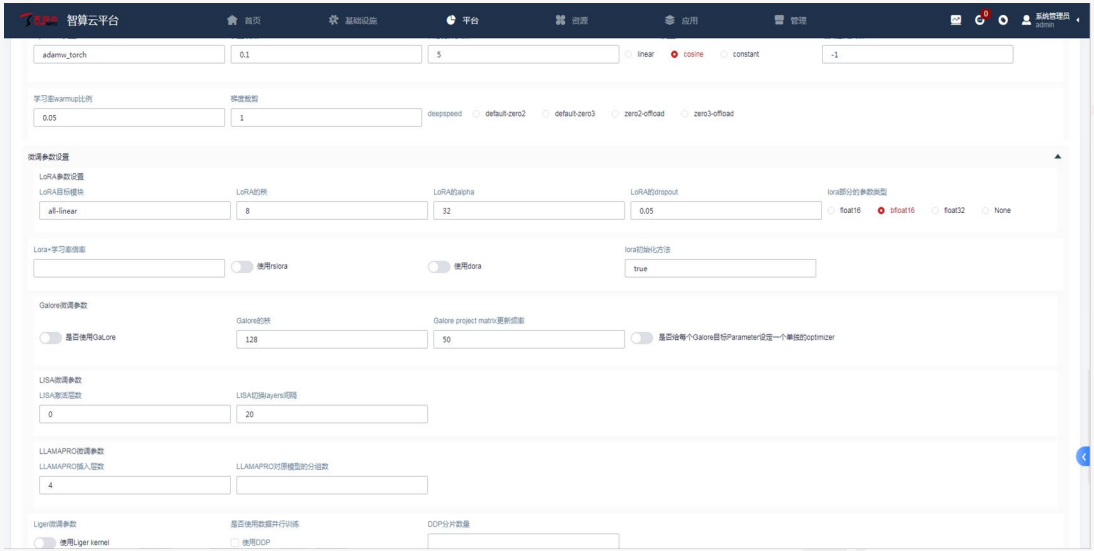
- 在创建训练、推理、量化与评测任务时，只需在图形界面中选择相应的预置模型或数据集，即可实现“拿来即用”

内置了上百种评测数据集

- 能够对纯文本和多模态模型进行评测，为用户评判模型和训练结果的优劣提供全面且精准的依据，确保模型质量达到预期。

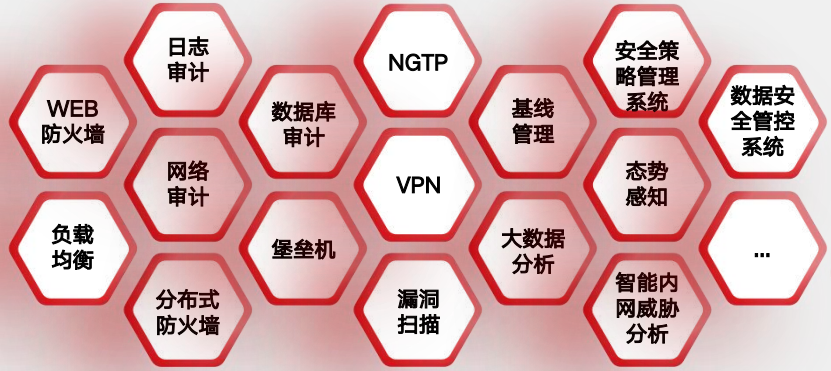
一站式AI开发

- **大模型全流程支持**：从数据处理、模型预训练、模型微调和模型推理等AI开发全流程支持，让算法工程师、业务架构师、应用开发者等可以专注、高效的完成业务工作
- **灵活模型选择**：可选预置基础/领域模型，第三方模型导入
- **端到端工作流**：一站式工作链，训练过程/结果可视化



基于模型应用的内生安全体系

一键解锁安全能力，为业务和DeepSeek等AI大模型提供全生命周期的安全保护



应用场景：「教育」一朵云，全场景支持

业务承载

- 一卡通
- 教务管理
- 人事管理
- 学生信息管理
- 后勤管理
- 财务管理
- 就业管理
- ...

超融合

办公/智慧教室/教学...

桌面云

AI 教学/科研

- 专业科研领域知识问答
- 科研资料检索
- 论文查询
- AI 教学
- ...

智算云

安全智算云平台

安全领域如何使用人工智能？

人工智能方向：机器学习、自然语言处理、认知推理、机器视觉、游戏道德、机器人等。

应用类型

- **预测：** 基于历史数据进行预测和分类
- **分析：** 从数据中分析、挖掘、发现隐藏的信息
- **抽取：** 从数据中抽取、提取、提炼特定的内容和上下文关系
- **生成：** 根据目标、应用、场景去生成特定的内容
- **决策：** 智能化的生成并去执行一个特定的动作或计划

基础要素

- **场景：** 识别分类、图像识别、语言翻译等
- **知识：** 领域专家、算法工程师、 AI 专家等
- **数据：** 海量、标注、高质量、高价值等
- **算力：** 云、服务器、IoT设备等
- **算法：** 机器学习、神经网络、Transformer架构等

安全有什么

■ **场景**
- 威胁检测、样本检测、态势分析、漏洞挖掘 等

■ **数据**
- 类型：流量、日志、样本、漏洞、代码、文本等
- **挑战：*海量？标注？高质量？高价值？***

■ **人/知识**
- 渗透测试工程师、运维服务工程师、溯源分析工程师、风险评估工程师

AI 安全模型 = 场景+知识+数据+算法+算力

大模型在安全领域内的应用，为网安提质增效

面向全能力的智能协同

能力跃迁

应用

检测增强 | 交付提升 | 辅助安全运营 | 加速互联互通

能力

提质

- 预测：恶意文件、样本分析、异常流量、威胁检测
- 分析：安全态势感知、异常行为分析
- 抽取：安全知识生成、文件指纹抽取

增效

- 生成：对抗样本生成、暴力破解密码
- 决策：智能策略下发、智能安全响应

算法

自然语言处理	抽取能力提升	大语言模型	通用环境（平台级）
深度学习	检测分析能力提升	小语言模型	资源受限环境（设备级）
机器学习		超微语言模型	资源受限环境（终端级）

生成、决策人机协同能力提升

算力

高性能算力输出（Kunpeng、Ascend） | 性能效率优化提升 | 麒麟软件 KYLINSOFT、OpenEuler | 集成微隔离+12类安全网元

数据

原始数据：样本、流量、日志、文本、…

加工后数据：政策、标准、方案、情报、安全知识、…

大模型应用场景-安全运营

大模型应用在网络安全运营中提升了防御效率，减少误报，增强了对复杂威胁的预测和响应能力。

- **威胁检测与分析：**

  通过分析庞大且复杂的数据集，识别潜在的安全威胁。通过机器学习，算法可不断学习最新的威胁模式，并对不正常的行为和异常流量模式进行实时警报。

- **行为分析：**

  用户和实体行为分析(UEBA)技术能够准确识别内部威胁和被攻破的账户。分析行为数据，从而可以发现与常规用户活动不匹配的行为，及时警告可能的内部威胁。

- **安全自动化（SOAR）：**

  大模型结合安全运维自动化(SOAR)。可以帮助自动化执行日常任务，如排序警报、响应低级威胁，并在需要时提供关键信息，支持快速决策。

- **漏洞管理：**

  大模型知识型图谱能力可以帮助漏洞管理制定主动的安全姿态，通过持续的自学习，扫描并预测新出现的软件漏洞和系统弱点，从而允许组织提前做好准备，防止潜在攻击。

- **威胁情报：**

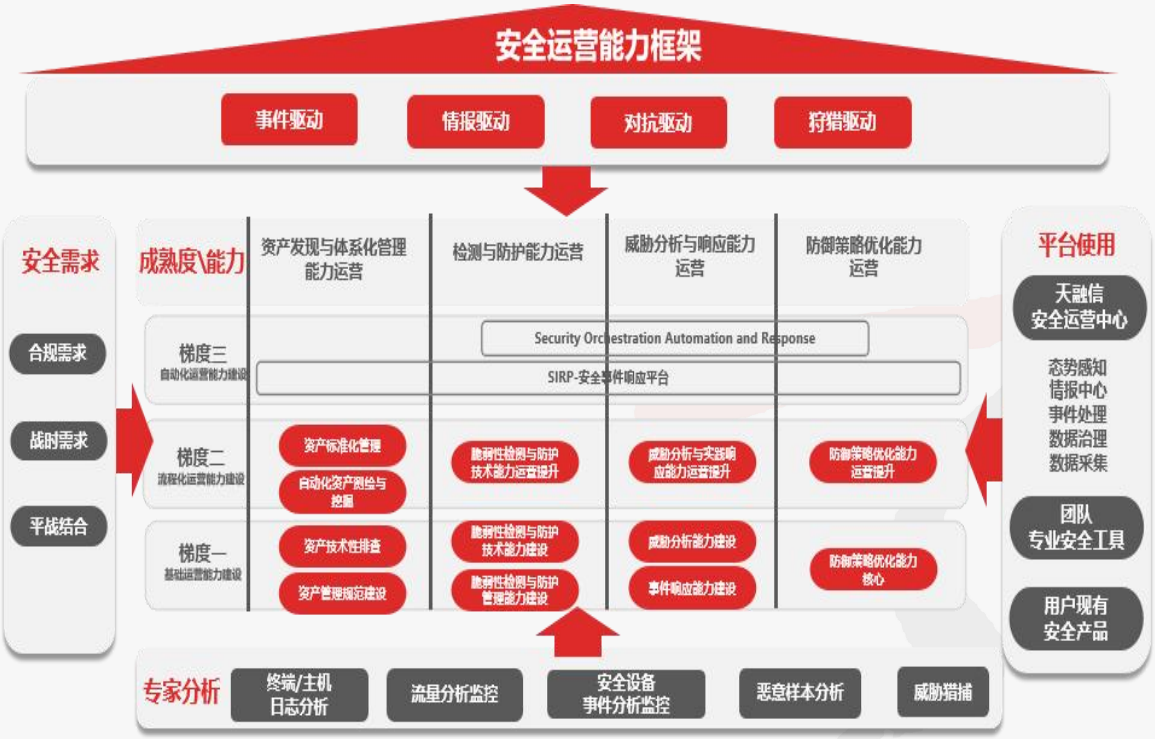
  利用大模型算法对大量的数据进行分析，识别并预测全球安全威胁的发展趋势。通过这种方式，组织能够更好地理解威胁环境，并据此制定防御策略。

- **网络安全评分：**

  综合分析多个安全维度，包括网络行为、密码策略、历史漏洞等，为企业提供一个动态的网络安全评分。这有助于企业及时了解自身的安全状况，并作出改进。

- **钓鱼监测：**

  对于各类邮件及附件、二维码、链接等各类信息邮件，可以实现快速监测、聚类及告警。



大模型应用场景-自动化运维【HW行动】

护网行动

每年护网都需要投入大量的人力物力，同时这种被动的防御方式对人员综合素质能力及设备功能、检测能力提出了原来越高的有效，当前护网对于一个安全事件随着攻击手段的升级并不限于可视化页面及固定功能组件的机械操作，对于代码及范式、shell脚本、API调用等能力要求越来越高，**而大模型AI的在护网的应用可以有效帮助护网技术人员较少上述护网技术压力，并提供有效代码、shell编写建议，包括但不限于以下能力：**

- **自然语言处理：**用户可以通过自然语言描述需求，将自然语言指令转化为可执行代码块，减少手动编码。
- **自动化脚本创建：**AI可以辅助编写用于系统维护、程序部署和任务自动化的shell脚本、任务识别转换Shell
- **脚本优化：**对现有脚本进行分析，并提供性能优化的建议。
- **任务识别转换：**用户只需要描述他们的自动化需求，大模型可以将这些需求转换成具体的shell脚本代码。
- **错误检测和修复：**检测脚本中的错误，并给出修复建议或自动修复。
- **历史数据聚类及快速降噪：**分析历史安全事件，帮助理解攻击者的行为和攻击模式。进行日志快速聚类分析，利用算法降低日志噪声。
- **实时监控：**通过实时分析日志数据，AI能够及时发现安全威胁，并触发警报。
- **模式识别：**AI可以从海量安全日志中识别异常活动的模式，以便快速定位问题。

备战阶段

全面发现检测、防护短板

全面检测现有防护体系、安全监测能力覆盖度。

全面补齐检测、防护短板

补齐安全检测及防护能力，梳理及优化防护策略。

决战阶段

7*24小时监控

对安全设备异常流量及告警实时监控。

研判处置及溯源

开展安全威胁研判，针对攻击行为进行溯源，绘制攻击者画像。

威胁情报共享

依托威胁情报共享机制，同步攻击特征，统一下发及处置。

总结阶段

总结复盘

复盘已发现的攻击行为，发现薄弱环节，固化有效策略，进一步完善安全防护体系。

大模型应用场景-数据安全

数据分类分级

过去：部分自动化工具、人工打标签、Python脚本

大模型：NLP自动化分类分级、敏感数据识别、自动化数据标记、上下文理解语义判定、合规判定

数据安全风险评估

过去：人工介入、设备检测为主，效率低、设备误判高

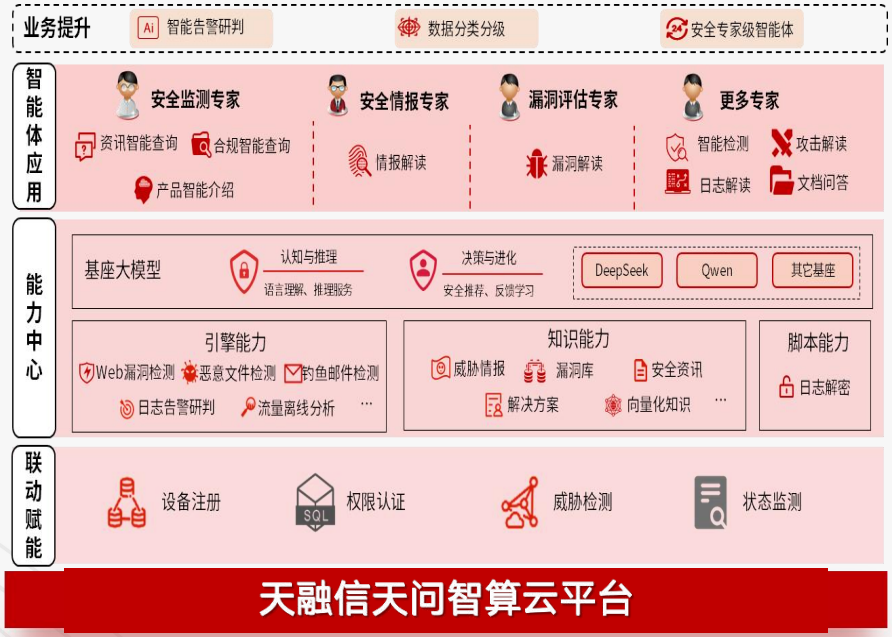
大模型：NL数据异常风险检测、数据流动风险趋势分析、风险持续量化、数据管控策略建议、数据风险合规检查等。

其他数据安全类

数据安全治理流程、数据分析洞察、数据整合、数据质量管理等

天问安全模型，赋能整体安全能力提升

天问大模型系统：基于自然语言交互方式提供智能问答、情报解读、漏洞解读、攻击解读等能力。并通过服务接口为天融信大数据分析系统、安全设备、多种平台联动赋能，构建更为智能化的安全运营防护体系。



态势感知

1.智能研判提效：大模型自动解析HTTP等复杂告警，分析、研判效率提升。

2.决策增强：可信度评分可监督五包，证据收集实现更完整报告的输出。

XDR

1.响应加速：大模型驱动SOAR 剧本生成与优化，快速封堵攻击，提升阻断精准度。
2.自动化攻防闭环：自动生成漏洞修复方案及验证脚本，分析数据推荐最小化修复策略。
3. 自适应防御：持续学习优化检测规则与响应逻辑，结合云端情报实时更新本地策略。
4.自然语言交互生成策略，自动生成多维度安全报告。

数据分类分级

1.降本增效：大模型自动完成数据识别、分类与分级，降低人工标注与审核成本。

2.精准合规：基于语义理解与上下文分析，提升分类准确率，确保符合相关法律法规和行业要求。

策略下发

1. 简化操作：通过自然语言对话生成策略，无需技术专家介入，降低分支机构运维门槛。
2. 策略一致性：总部统一编排下发策略，规避不同分支人工配置差异导致的安全风险。
3. 风险预控：自动校验策略逻辑（如冲突检测、权限最小化），防止配置错误或无效。

每一个新技术的发展与应用，都会带来全新的安全风险激增

AI 应用引入更多安全风险

AI 加持的敌手更强大

网络攻击

代码生成

深度伪造

……

人工智能安全风险

OWASP LLM应用十大风险（TOP10）

2023	2024
LLM01 提示注入	LLM01 提示注入
LLM02 敏感信息批露	LLM03 供应链
LLM03 供应链	LLM02 敏感信息批露
LLM04 数据和模型投毒	LLM06 过度代理权限
LLM05 不当输出处理	LLM05 不当输出处理
更 LLM06 过度代理权限	LLM04 数据和模型投毒
新 LLM07 系统提示泄露	LLM10 无界消费
新 LLM08 向量和嵌入的弱点	LLM09 虚假信息
更 LLM09 虚假信息	新 LLM08 向量和嵌入的弱点
更 LLM10 无界消费	新 LLM07 系统提示泄露

人工智能内生安全风险

模型算法安全风险

1)可解释性差
2)偏见、歧视
3)鲁棒性弱
4)被窃取、篡改
5)输出不可靠
6)对抗攻击

数据安全风险

1)违规收集使用数据
2)数据“投毒”
3)数据标注不规范
4)数据泄露

系统安全风险

1)缺陷、后门被利用
2)算力安全风险
3)供应链安全风险

人工智能应用安全风险

网络域安全风险

1)信息内容
2)混淆事实、误导用户绕过鉴权
3)不当使用引发信息泄露
4)滥用于网络攻击
5)模型复用缺陷传导

现实域安全风险

1)诱发传统经济社会安全
2)用于违法犯罪活动
3)两用物项和技术滥用

认知域安全风险

1)加剧“信息茧房”效应
2)用于开展认知战

伦理域安全风险

1)加剧社会歧视偏见、扩大智能鸿沟
2)挑战传统社会秩序
3)未来脱离控制

AI大模型安全防护技术措施

模型安全服务

- AI管理咨询服务
- 数据安全服务
- 模型安全服务
- 供应链安全服务
- 模型合规服务
- 网络安全服务

API/数据泄漏监测

隐私泄露监测 | 接口风险监测 | 恶意行为监测 | 权限滥用监测 | 输出内容监测 | 流量异常监测

模型应用

业务应用1 | 业务应用2 | 业务应用3 | 业务应用4 | 业务应用5 | ……

AI大模型安全防护系统

敏感词防护 | 反向诱导防护 | 提示泄露防护 | 越狱防护 | 指令劫持防护 | 开发者防护 | 弱语义防护

模型输出

语义理解 | 数据分析 | 逻辑推理 | 预测分析 | 辅助决策 | 事件研判

模型训练

训练数据集	模型训练/优化	模型组件集	模型安全测评	模型脆弱性安全
公开/自有数据集	监督/半监督训练	模型架构集成	数据异常测评	模型漏洞检测
数据标注	算法优化/剪切	神经网络层	训练合规测评	模型基线核查
结构化/非结构化	超参数调整	损失函数	对话机制测评	模型威胁检测

模型部署

云主机 | 主机安全 | 容器 | 容器安全

基础设施

GPU/CPU集群 | 交换/网络设备 | 存储集群 | 服务器集群 | 通用网络安全设备

模型安全运营

- 模型备案管理
- 安全防护管控管理
- 安全事件响应处置
- 安全风险管控管理
- 检测问题处置稽核

模型安全防护提升大模型生命周期安全

AI安全防护管控

策略管理
- 黑白名单策略
- 敏感词策略
- 身份认证与授权
- ……

防御管理
- 防御任务运营
- 防御操作管理
- 防御任务分析
- ……

基线管理
- 镜像基线管理
- 主机基线管理
- 容器基线管理
- ……

运行状态监测
- 容器运行状态
- 接口调用监测
- 访问频率监测
- ……

文件活动 / 安全事件 / 告警信息 ↕ 策略下发

容器信息 / 镜像漏洞 / 进程活动 ↕ 策略下发

流量信息 / 访问信息 / 认证和鉴权 ↕ 策略下发

策略匹配 / 行为日志 / 内容安全事件 ↕ 策略下发

安全防护

主机安全
- 恶意代码检测
- 策略删除
- 主机外联检测
- ……

容器安全
- 微隔离
- 镜像准入
- 容器逃逸检测
- ……

API数据防泄漏系统
- 流量管控
- 身份认证
- 访问控制
- ……

AI大模型安全防护系统
- 内容审核
- 敏感词过滤
- 提示词注入检测
- ……

防护对象

物理服务器 | 操作系统 | 云主机 | 容器 | API接口 | 数据安全 | 大模型

模型安全评估验证构建合规的大模型能力

模型安全运营中心

入网三同步流程管理

模型开发检测

模型入网前检测

模型运行检测

模型安全检测任务

基础环境检测任务管理

任务下发

- 网络安全检测任务
- 数据安全检测任务
- 主机安全检测任务
- 系统安全检测任务

输出

- 脆弱性检测报告
- 基线核查报告
- 软件成本分析报告

AI模型检测任务管理

任务下发

- 语义理解异常测评
- 数据异常测评
- 对话机制异常测评
- 系统异常测评

输出

- 模型安全评分
- 模型风险清单

检测问题稽核

运营中心采集安全检测问题

问题分析研判

问题处置

工单处置　联动处置

问题复核

一线 监控角色　二线 专家角色　三线 研发角色

任务下发　检测结果

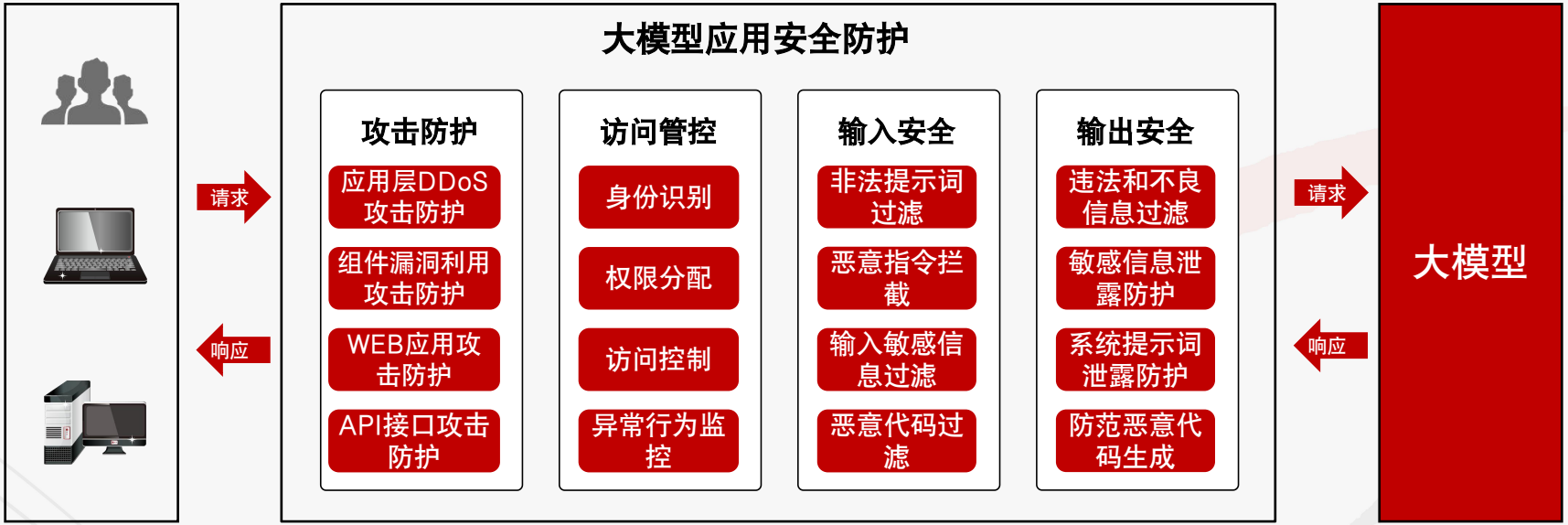
传统安全检测设备

任务下发　检测结果

AI安全测评系统

大模型应用安全防护能力加强模型应用保障能力

专注为企业级大模型提供应用安全防护，可针对大语言模型所面临的特有安全威胁（如应用层DDoS攻击、供应链漏洞攻击、提示词注入攻击、恶意tokens消耗等）进行实时检测和防御，并提供模型访问安全管控、不安全输出过滤、敏感信息防泄露等能力。



API数据安全监测能力提供接口调用可监测可审计

- 面向大模型API服务的数据安全威胁，构建『四维智能监测中枢』，实现从协议层异常调用拦截、内容层敏感信息过滤、资源层高频滥用防控到模型层数据泄露监测的全栈式数据安全监测手段，实现对大模型应用服务的可监测、可审计、可审计。

接口行为风险监测

- 异常API调用模式识别
- 非授权功能接口调用监测
- 协议规范符合性识别

模型违规行为监测

- 模型参数窃取行为监测
- 中间层数据缓存监控
- 模型组件的漏洞泄露监测

模型内容合规检测

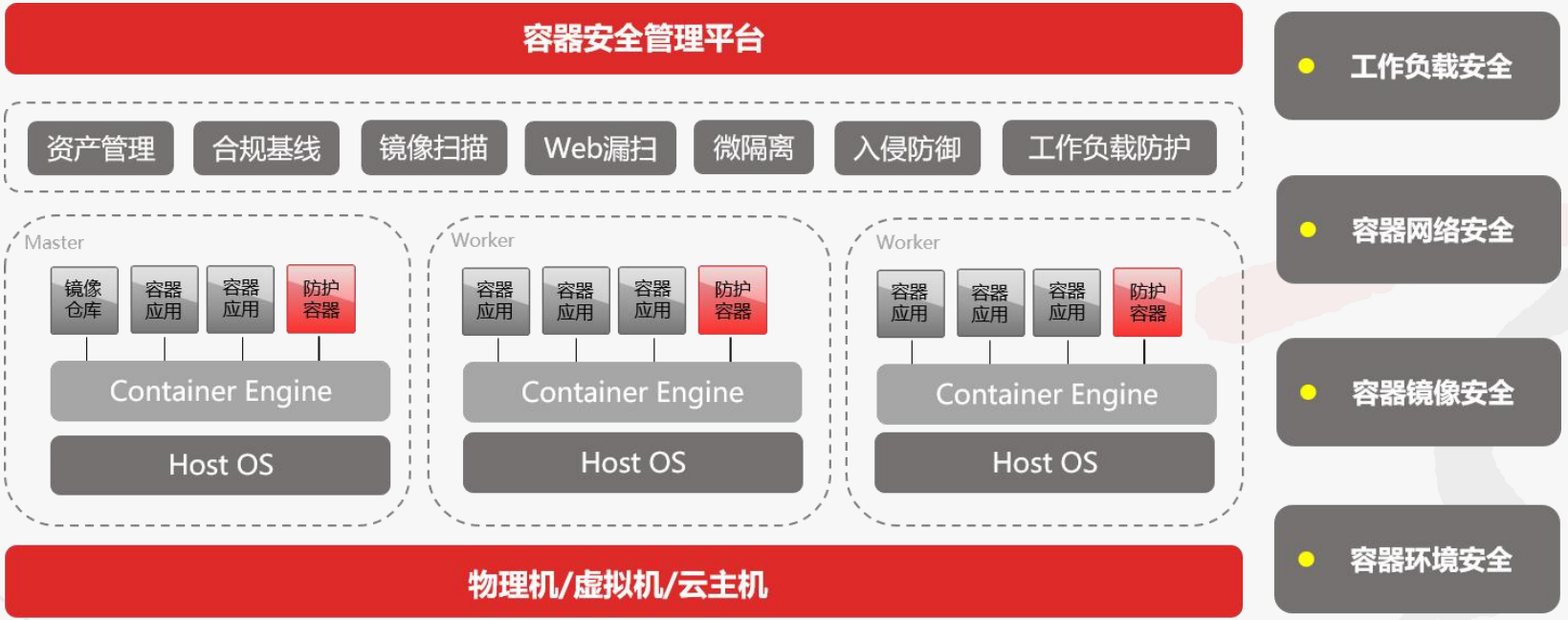
- 敏感信息传输检测
- 政治/伦理/法律合规审查
- 多模态内容深度解析

模型资源滥用监测

- 高频访问DDoS特征识别
- Prompt注入攻击监测
- 模型推理资源配额基线监测

容器安全防护能力提升基础计算环境安全

容器安全防护能力以容器环境安全、容器镜像安全、容器网络安全、工作负载安全四个维度为切入点，帮助解决大数据模型部署在容器后存在的容器逃逸攻击、供应链攻击等安全风险。

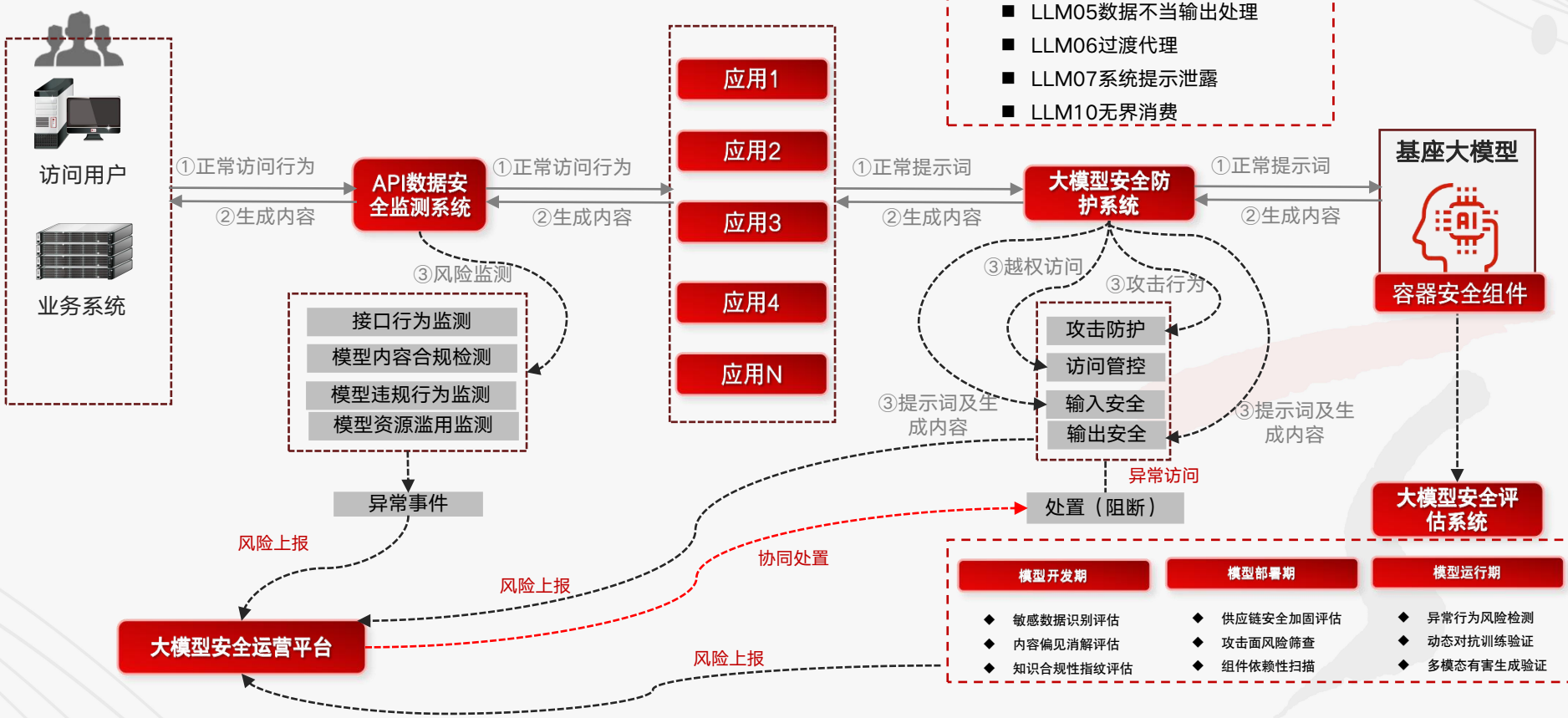


AI大模型安全防护实践

解决大模型风险点

- LLM01提示词注入风险
- LLM02敏感信息泄露
- LLM05数据不当输出处理
- LLM06过渡代理
- LLM07系统提示泄露
- LLM10无界消费

天融信 TOPSEC
证券代码：002212



做有实力的网安垂域大模型

天融信 TOPSEC
证券代码：002212

网信办大模型备案

境内深度合成服务算法备案清单（2024 年 6 月）

序号	算法名称	角色	主体名称	应用产品	主要用途	备案编号
61	天融信天问安全文本生成算法	服务提供者	北京天融信网络安全技术有限公司	[illegible]	应用于文本生成场景，根据用户输入的文本，生成网络安全领域相关的文本。	网信算备110108071863901240019号
62	天融信天问安全文本生成服务算法	服务技术支持者	北京天融信网络安全技术有限公司	—	应用于文本生成场景，服务于企业端客户，根据用户输入的文本，生成网络安全领域相关的文本。	网信算备110108071863901240027号

境内深度合成服务算法备案

北京市生成式人工智能服务新增已备案信息

序号	大模型名称	服务提供者	上线备案编号
5	天问	北京天融信网络安全技术有限公司	Beijing-TianWen-202409250036

生成式人工智能服务备案

第三方认证

能力评测
首批

大模型安全
服务能力评定
工信部软件测评中心

安全性评测
首批·A级认定

大模型产品
安全性检测
工信部软件测评中心

能力评估
基础网络安全能力

安全大模型基础
网络安全能力评估
中国信息通信研究院

知识成果
自主创新

软件著作权
国家版权局

标准专利参与

参与AI相关标准编制11项

- TC260-003《生成式人工智能服务安全基本要求》
- 网络安全领域独家参编《人工智能安全标准化白皮书（2023 版）》
- 《人工智能产品、应用及服务供应基础安全能力评估方法》
- …

AI相关专利申请279项，已授权51项

- 一种流量的监控方法、模型的训练方法、装置及存储介质
- 威胁行为检测和模型建立方法、装置、电子设备及存储介质
- 一种DGA域名检测方法及装置
- …

主要荣誉

- 【IDC】网络安全领域大模型代表厂商、大模型能力获重点推荐
- 【CCF】《面向网络安全关系抽取的大型语言模型数据增强》获“第39次全国计算机安全学术交流会”唯一优秀论文奖
- 【CCIA】天问大模型荣获“2024年大模型安全实践优秀案例”
- 【CSA】《天融信基于AI赋能的安全运营解决方案》获 CSA 2024安全磐石奖…

谢谢！