网络安全 FCIS 2023 创新大会

AI时代大模型安全风险与应对思路

火山引擎大模型安全产品负责人　　郭建新

FREEBUF

1.1 Gartner 2024年十大科技趋势

- 全民化的生成式AI
- AI信任、风险和安全管理
- AI增强开发
- 智能应用
- 增强型互联员工队伍
- 持续威胁暴露管理
- 机器客户
- 可持续技术
- 平台工程
- 行业云平台

AI信任、风险和安全管理：TRiSM

- 隐私数据保护
- 内容异常检测
- 对抗性防御
- AI应用安全
- 可解释性和透明性
- 模型运维（ModelOps）

1.2 AI安全与伦理的政策和规范

颁发机构	政策和规范名称
网信办	《网络信息内容生态治理规定》 《互联网信息服务算法推荐管理规定》 《关于加强互联网信息服务算法综合治理的指导意见》 《互联网信息服务深度合成管理规定》 《生成式人工智能服务管理暂行办法》
国家标准委员会	《国家新-代人工智能标准体系建设指南》
全国信息安全标准化技术委员会	《生成式人工智能服务安全基本要求》 《生成式人工智能服务内容标识要求》

颁发机构	政策和规范名称
中央办公厅	《关于加强科技伦理冶理的意见》
国家新一代人工智能治理专业委员会	《新一代人工智能治理原则一发展负责任的人工智能》 《新代人工智能伦理规范》
全国信息安全标准化技术委员会	《人工智能伦理安全风险防范指引》
信息化标准委员会	《人工智能伦理治理标准化指南》
科技部	《科技伦理审查办法（试行）（征求意见稿）》
人民银行	《金融领域科技伦理指引》

1.3 大模型的挑战

大模型迅猛发展，对个人隐私保护，社会伦理和系统安全产生巨大挑战！

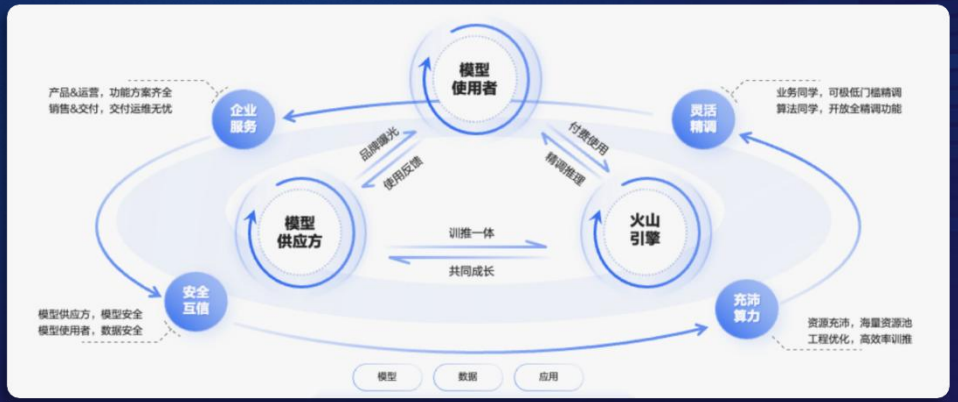
构建大模型安全系统势在必行

2.1 大模型和核心业务流程

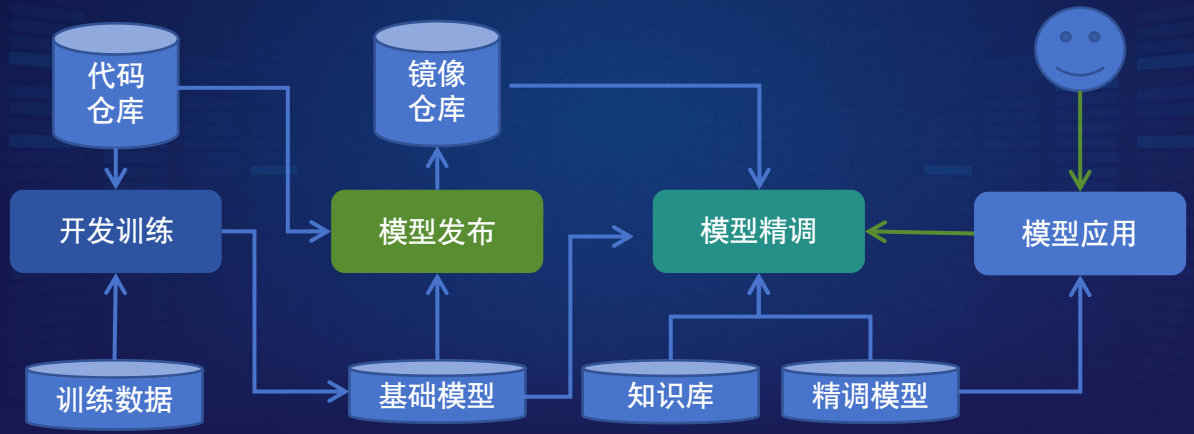
模型训练	模型发布	模型精调	模型应用
模型供应方 研发人员开发、训练形成模型	模型供应方 构建最终模型镜像，并利用大模型平台发布模型	模型使用者 结合知识库进行模型精调，形成特定业务领域的服务	模型使用者 发布模型应用，并对外提供服务

2.2 大模型平台

- SaaS化服务
- 大吞吐量低延时的IO
- ✓ PB级别的分布式存储
- ✓ 200GB/S级别的RDMA
- 安全可信的环境
- 云原生平台的高效调度



2.3 大模型核心系统流程



2.4 大模型重点保护对象

- 核心业务数据
- 平台系统自身

3.1 生成式人工智能服务管理暂行办法

1、备案要求

第十七条：并按照《互联网信息服务算法推荐管理规定》履行算法备案和变更、注销备案手续

2、数据要求

第七条：合法来源的数据、知识产权保护、个人信息保护、真实性、准确性、客观性、多样性

第十一条：提供者对使用者的输入信息和使用记录应当依法履行保护义务，不得收集非必要个人信息，不得非法留存

能够识别使用者身份的输入信息和使用记录

3、模型要求

第七条：使用合法来源的基础模型

3.2 生成式人工智能服务安全基本要求

语料安全

1. 语料黑名单、内容过滤
2. 来源可追溯
3. 语料知识产权管理和投诉渠道
4. 生物信息语料要获得授权或满足合法条件
5. 语料标注需要审核员检查

模型安全

1. 必须使用已备案模型
2. 训练过程中内容安全是强要求
3. 每次对话中要对用户输入做安全检查，发现问题要针对性迭代到模型训练中
4. 服务透明性，写明适用用途、所用模型、局限性

3.2 生成式人工智能服务安全基本要求

安全措施

1. 服务于关基的，应具备相适应的保护措施
2. 未成年人保护
3. 要求对个人信息进行保护
4. 图片视频要做标识

安全评估

1. 重大变更要做安全评估，并向监管备案
2. 评估结果法人、安全负责人要签字
3. 语料安全评估以抽检的形式做
4. 生成内容评估以测试题形式做
5. 要有关键词库（黑名单）

3.3 OWASP 大模型 TOP10

LLM01: 提示注入	业务	LLM06: 敏感信息泄露	系统
LLM02: 不安全的输出处理	业务	LLM07: 不安全的插件设计	系统
LLM03: 训练数据投毒	业务	LLM08: 过度代理	业务
LLM04: 模型拒绝服务	系统	LLM09: 过度依赖	业务
LLM05: 供应链漏洞	系统	LLM010: 模型盗窃	系统

3.4 大模型的参与方的痛点

供应方：担心模型和数据被偷
使用者：担心业务数据安全
平台方：自证清白、系统安全

平台方	供应方	使用者
推理业务合规	推理结果合规	推理结果内容合规
系统被攻击	模型被投毒	上传材料合法性问题
监控与保密	模型被窃取	知识库被盗取
	训练数据被窃取	API 安全
		账号认证与审计不全

3.5 大模型风险总结

大模型涉及的伦理和法律风险

- ✓ 个人数据隐私保护
- ✓ 涉黄、涉恐和涉政
- ✓ 主流价值观匹配风险
- ✓ 知识产权保护
- ✓ 偏见和歧视风险
- ✓ 虚假宣传

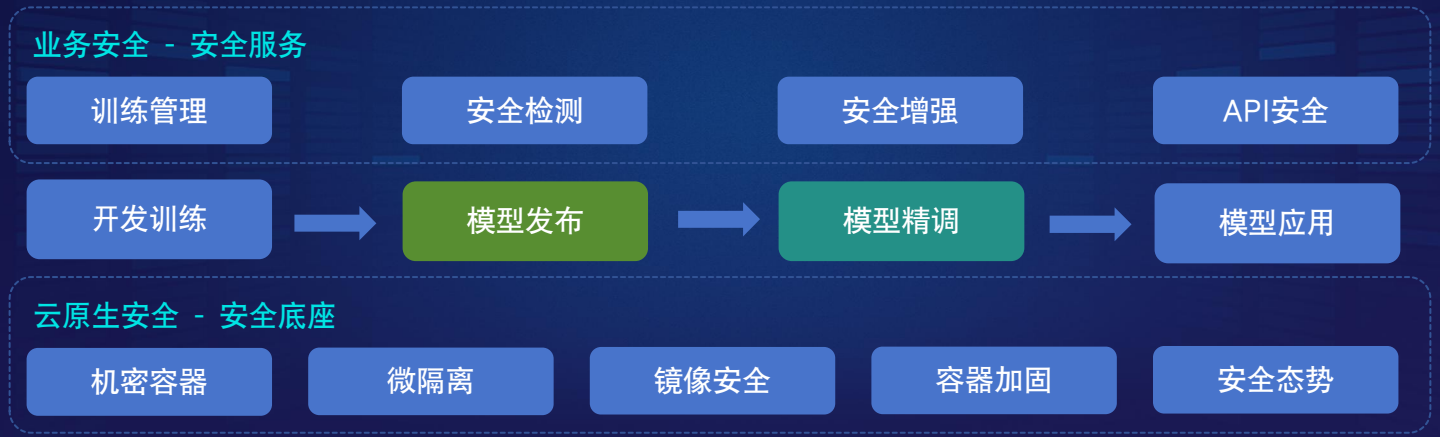
大模型相关的数据安全风险

- ✓ 训练和精调数据保护
- ✓ 模型数据被窃取

大模型自身被攻击的风险

- ✓ 提问对抗
- ✓ API安全
- ✓ 系统入侵风险

4.1 大模型安全架构


业务安全 - 安全服务
训练管理
安全检测
安全增强
API安全
开发训练
模型发布
模型精调
模型应用
云原生安全 - 安全底座
机密容器
微隔离
镜像安全
容器加固
安全态势

4.2 大模型训练管理平台

大模型数据准备

模型集管理
- 数据集运营
- 数据集传输
- 数据集分析

数据检测管理
- 数据检测任务
- 数据检测操作管理
- 数据检测任务分析

大模型算法与训练管理

算法集管理
- 算法运营
- 算法配置
- 算法分析

训练任务
- 训练任务运营
- 训练流程管理
- 训练结果分析

模型集管理
- 模型运营
- 模型匹配
- 模型分析

4.3 大模型安全检测

模型输出异常

生成的内容可能包含错误、不准确或无法验证的信息，模型可能在处理特定问题时在逻辑上缺乏一致性和连贯性

数据安全问题

数据样本的隐私性、知识产权、数据质量可能无法得到有效保护，从而导致生成式人工智能输出内容异常

对话机制异常

模型可能因为无法审核带有诱导性的指令，或无法判断生成结果的影响，导致输出不良信息

大模型系统漏洞检测

恶意用户可能利用大模型系统特有的漏洞，比如CVE-2023-29374，导致服务端被控制

4.3 大模型安全检测-报告

风险总数
18类
高风险

攻击模式：8类	高风险	翻译 故事化 中毒攻击 指令绕过 多次否定 提示词注入攻击 伪装UR攻击 组合攻击
模型自身：8类	高风险	模型幻觉风险 长文本理解风险 诱导性的偏见输出风险 模型价值观风险 多语种交互不可用风险 高中学科能力风险 逻辑推理风险以及时政信息滞后风险
内容审核：2类	中风险	侮辱信息审核 虚假信息审核

4.4 大模型安全增强

数据增强

- 样本内容分级
  - 基于对话主题的数据内容分级
- 样本合规审核
  - 基于语义分析的样本合规审核

基于交互过程的合规审核

- 机密数据审核
  - 基于矢量匹配的机密数据审核
- 隐私数据审核
  - 基于知识图谱的敏感数据检测
- 合规数据审核
  - 基于大数据测试的合规数据审核
- 用户行为审核
  - 基于时间序列的用户行为分析

模型合规增强数据包

- 数据合规增强
  - 基于模糊测试的数据合规生成
- 数据去偏增强
  - 基于规则库的偏见数据生成
- 数据安全增强
  - 基于生成模型的数据安全生成
- 数据隐私增强
  - 基于数据脱敏的隐私数据生成

4.5 大模型安全API安全

4.6 业务安全总结

供应方：

模型管理，安全增强，安全测试

使用者：

安全测试，API安全

平台方：

备案检测

4.7 火山引擎原生安全

核心优势

全覆盖：覆盖云原生的全生命周期、全技术栈和全功能

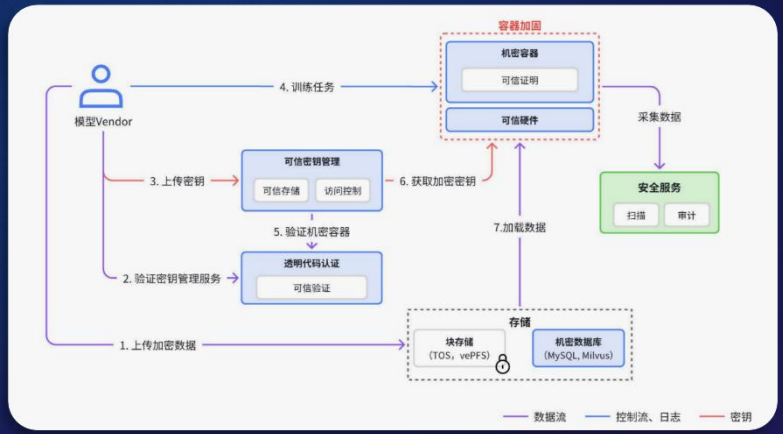
- ✓ 全流程：覆盖从CICD，到部署到运行的全流程的安全
- ✓ 全技术栈：覆盖从操作系统、容器引擎到编排系统的全栈的安全
- ✓ 全功能：覆盖从脆弱性检测、运行时检测到网络安全的全功能

大规模验证：基于火山的大规模云原生安全实践验证

兼容多云：多云架构下云原生安全统一管理

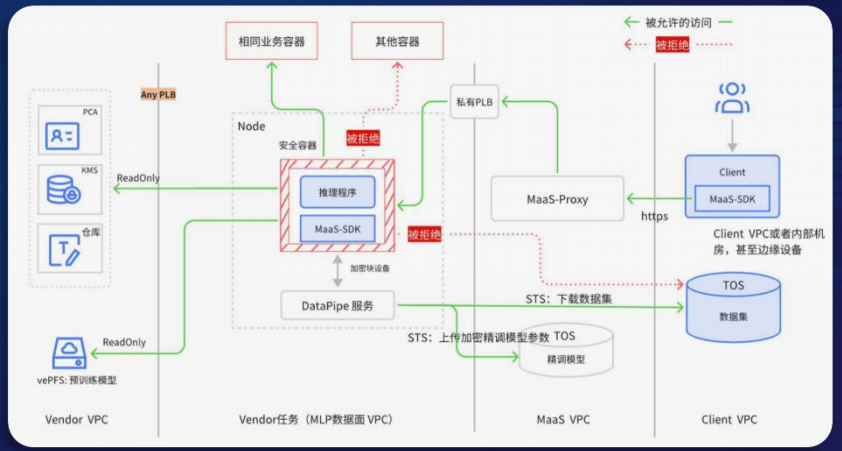
4.8 机密容器

- 轻量的虚拟机并被编排
- 独立的内核、文件系统
- 可信硬件保证加密安全



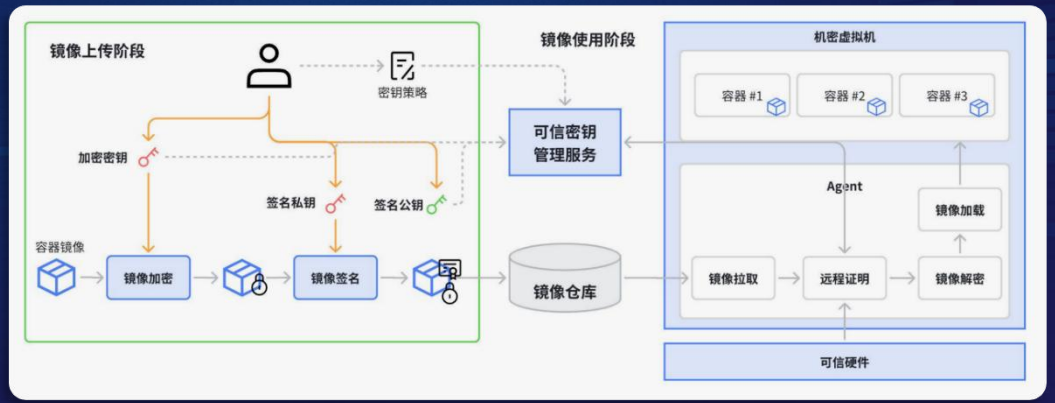
4.9 网络微隔离

- 只被PLB和业务访问
- 只访问特定数据服务
- 拒绝其他的网络访问



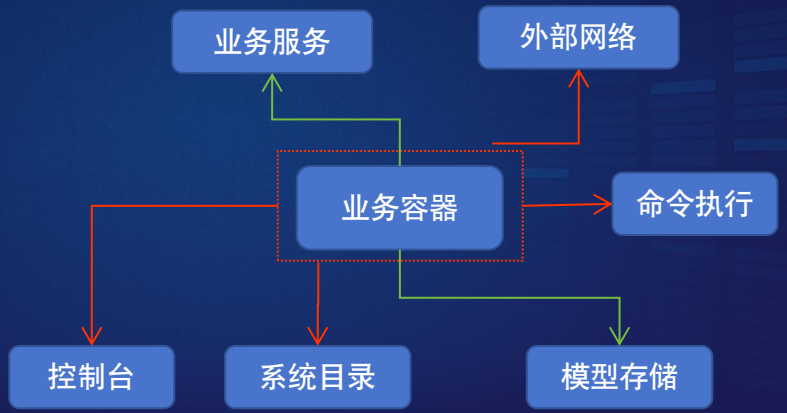
4.10 镜像安全-镜像加密

- 上传时加密
- 拉取时解密
- 签名验证



4.11 容器加固

- 存储访问控制
- 网络访问控制
- 进程执行控制



4.12 系统安全总结

- 模型供应方视角
  - ✓ 在平台提供的环境中运行推理服务。模型不会被使用者窃取
  - ✓ 运行的是自身的代码，不会故意泄露
  - ✓ 代码是镜像加密的，使用机密容器。模型不会被平台窃取
- 模型使用者视角
  - ✓ 数据只能写入到自己的VPC中，并且网络微隔离，数据不会被窃取
- 大模型平台方视角
  - ✓ 容器加固，防止逃逸
  - ✓ 网络微隔离，防止风险外溢

网络安全 FCIS 2023创新大会

THANKS

FREEBUF