腾讯安全 | 腾讯安全云鼎实验室

模型有界 安全无疆

腾讯安全沙龙第2期（西安站）

关于我
About me

@bayuncao

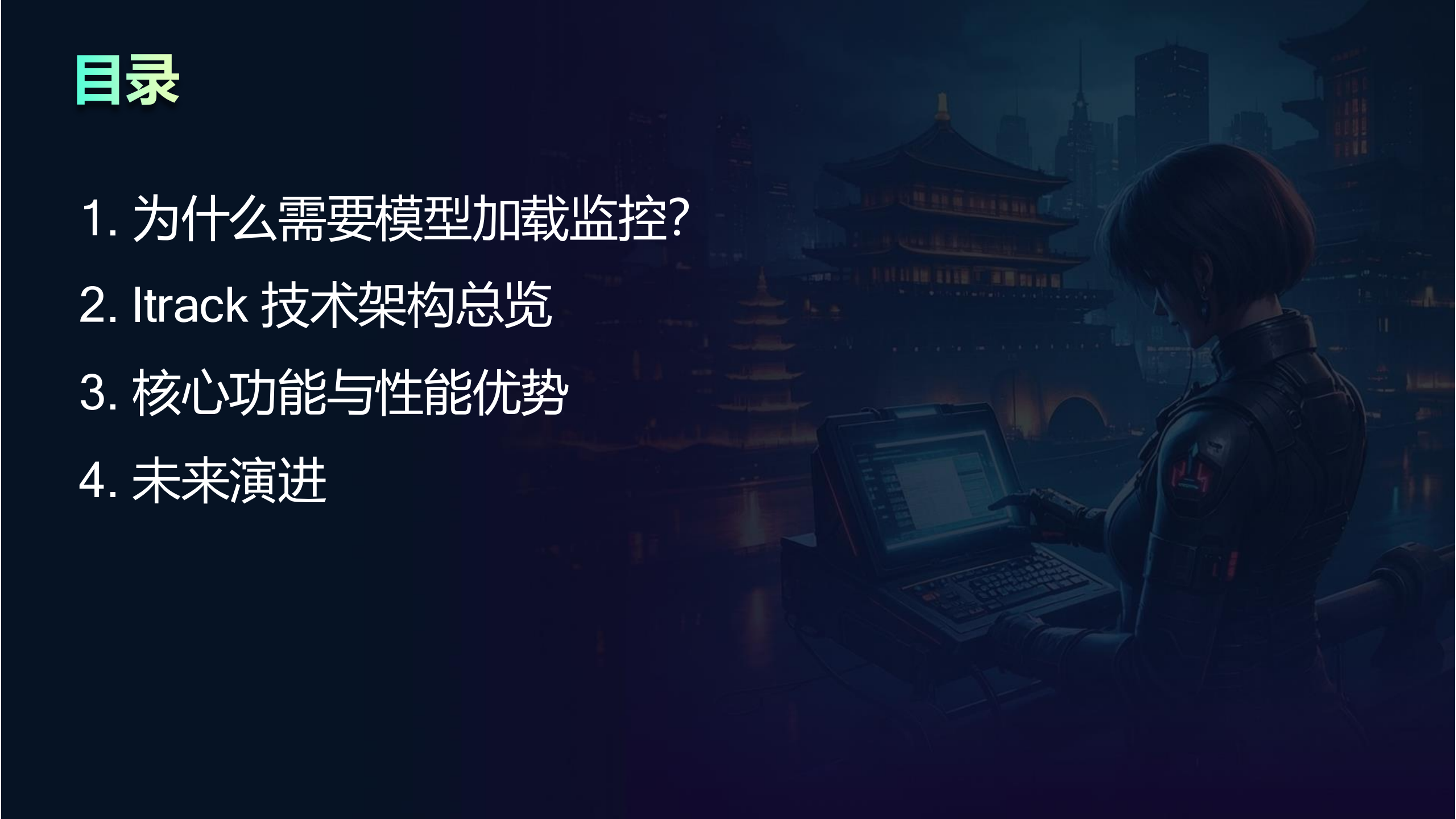
ChaMD5安全团队AI组负责人，专注于AI安全领域的研究者与实践者，在开源大模型漏洞挖掘方向取得多项突破性成果，持有多个CVE漏洞编号及通用漏洞证书，擅长将工程能力与安全研究结合，全栈开发，自研AI代码审计系统，持续深耕大模型供应链安全、越狱攻防及AI Agent漏洞自动化挖掘领域，致力于构建AI时代的新型防御体系。

Itrack - Security Observability Framework for ML/AI Model File Loading

让模型加载过程透明可溯

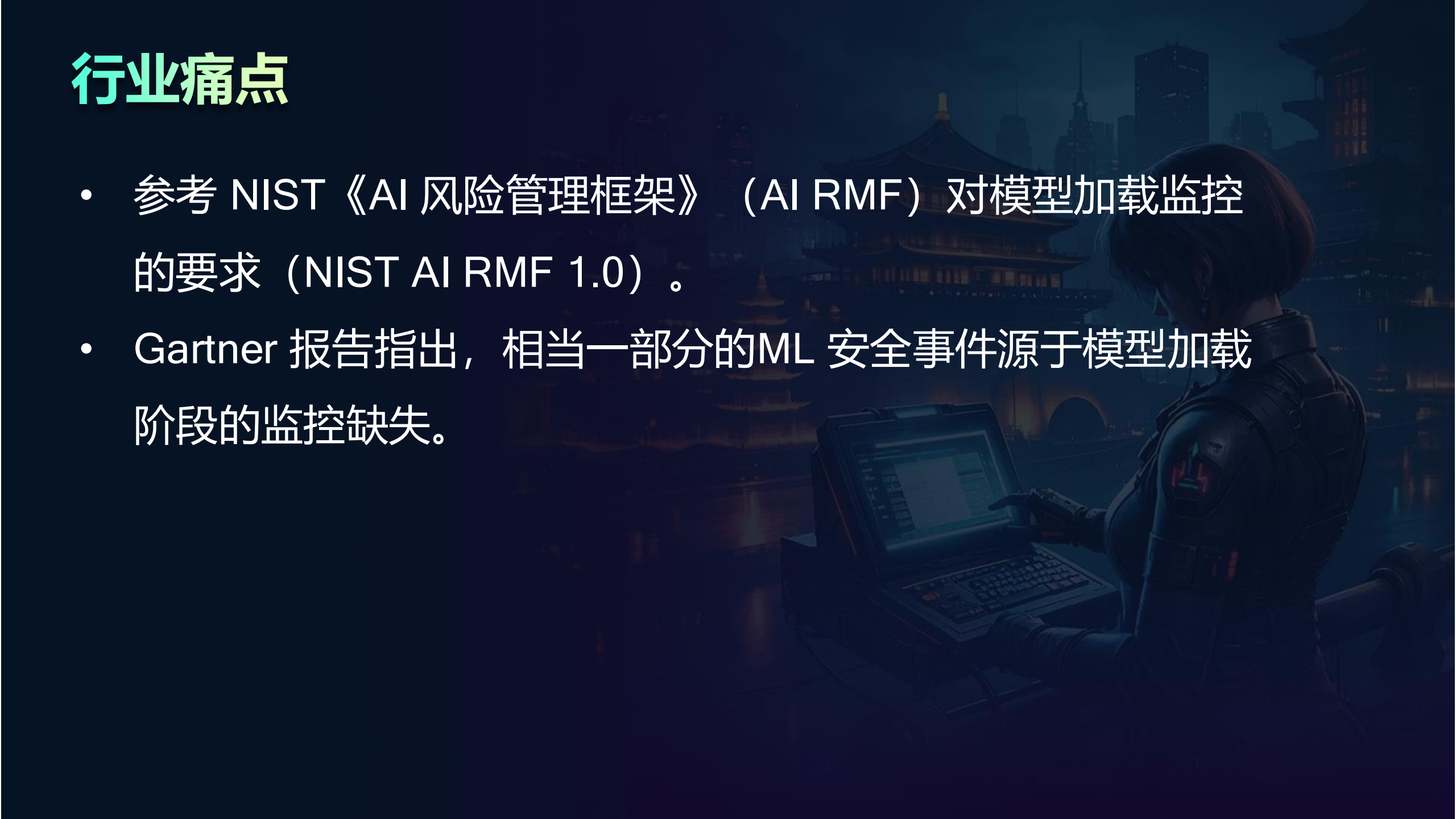

ltrack 正式发布
https://github.com/mxcrafts/ltrack

目录



行业痛点

- 参考 NIST《AI 风险管理框架》（AI RMF）对模型加载监控的要求（NIST AI RMF 1.0）。
- Gartner 报告指出，相当一部分的ML 安全事件源于模型加载阶段的监控缺失。



模型文件加载行为不可见（隐蔽性威胁）

- 2022年 PyTorch 恶意依赖项攻击，包名为 torchtriton，包含一个二进制文件，除了窃取主机名、DNS 配置、用户名、shell 环境等系统信息外，还会将*/etc/hosts*、*/etc/passwords*、*~/.gitconfig*、*~/.ssh /**的内容，以及在用户主目录中找到的前 1000 个文件上传到外部服务器。
- 根据《大模型可信赖研究报告》预训练数据集中可能包含来源不明或被恶意投毒的数据，若未严格检测，模型可能学习到有害信息并泄露隐私

传统监控工具无法感知 ML 上下文

- 上下文缺失：LLM 推理需加载多阶段子模型（如 Tokenizer、Embedding 层），传统工具无法标记各阶段资源消耗，导致无法检测 “模型加载劫持” 攻击（如恶意进程伪装成合法推理服务）。

- 动态行为盲区：Oracle云漏洞导致挖矿程序驻留。

合规审计缺乏细粒度日志

- 合规漏洞：2024 年OpenAI在2023年3月的数据泄露事件中未能及时通知监管机构，并且在没有合法依据的情况下使用用户数据训练ChatGPT，违反 GDPR 数据保护条例，被处罚1500万欧元。

- 审计实践：Google Cloud AI 安全指南要求记录模型加载的完整依赖链，包括临时文件、内存映射和子进程行为。

解决方案对比

方案类型	监控粒度	性能损耗	ML 上下文感知	技术原理	适用场景	局限性
传统 HIDS	进程级	高（>15% CPU）	❌	基于规则匹配系统调用（如 open、execve）	通用服务器安全监控	无法关联模型版本、依赖路径等业务语义
应用日志	业务级	低（<1% CPU）	部分	依赖 ML 框架自身日志（如 TensorFlow/PyTorch）	开发调试、基础行为审计	无法捕获底层依赖库加载、容器逃逸行为
🔥ltrack	系统级 + 进程级	<3% CPU	✅	eBPF 挂钩文件操作 + 用户态解析 ML 元数据	生产环境实时监控、合规审计	需 Linux 4.4 + 内核、需 CAP_BPF 权限部署

文件层攻击面（模型权重篡改）

- PyTorch 供应链攻击：2022年恶意 PyPI 包伪装成 PyTorch 扩展库，劫持模型加载流程。
- Hugging Face 模型投毒：Hugging Face披露部分用户上传的PyTorch模型文件（.bin）可能通过pickle反序列化执行恶意代码。

Itrack（Features）

- 监测模型文件加载时的inode变化与哈希值，关联加载进程的容器上下文（如 Kubernetes Pod ID）。

执行层攻击面（恶意依赖注入）

- PyPI 恶意包攻击：2024 年攻击者上传tensorflow-nightly-malicious包，窃取 GPU 算力。
- conda 仓库投毒：某金融科技公司的内部conda仓库遭入侵，攻击者上传同名但版本号更高的恶意opencv-python包。由于企业依赖解析策略缺陷，内部模型训练服务优先拉取了恶意版本。

Itrack（Features）

- 通过 eBPF 跟踪动态链接库加载事件（dlopen），标记非白名单依赖（如/tmp/libhack.so）。

网络层攻击面（模型泄露）

- 模型窃取攻击：攻击者通过 API 高频查询重构 LLM 参数。

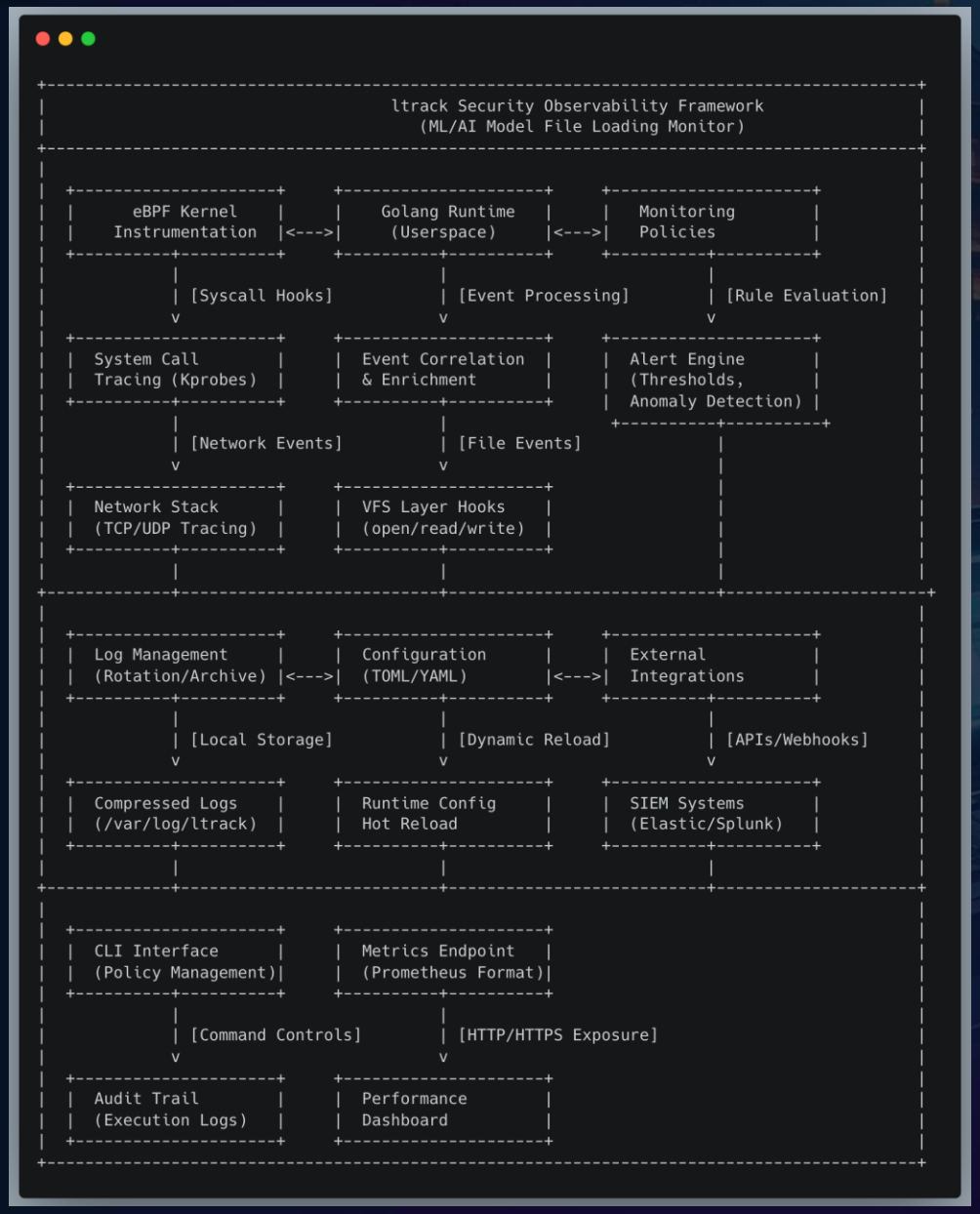
Itrack（Features）

- 捕获敏感文件（如*.pt）的sendfile系统调用，关联进程网络行为（如 TCP 连接目标 IP）。

ltrack 在 MITRE 框架中的覆盖范围

攻击面	覆盖战术	覆盖技术数	检测能力示例
文件层	Defense Evasion	2	文件哈希校验、进程签名验证
执行层	Persistence, Privilege Escalation	2	依赖路径监控、权限链分析
网络层	Exfiltration	2	文件传输行为关联、网络连接审计
总计	3 大类	6 项技术	覆盖 ML 全生命周期攻击链

ltrack 架构总览


ltrack Security Observability Framework
(ML/AI Model File Loading Monitor)
eBPF Kernel Instrumentation
Golang Runtime (Userspace)
Monitoring Policies
[Syscall Hooks]
[Event Processing]
[Rule Evaluation]
System Call Tracing (Kprobes)
Event Correlation & Enrichment
Alert Engine (Thresholds, Anomaly Detection)
[Network Events]
[File Events]
Network Stack (TCP/UDP Tracing)
VFS Layer Hooks (open/read/write)
Log Management (Rotation/Archive)
Configuration (TOML/YAML)
External Integrations
[Local Storage]
[Dynamic Reload]
[APIs/Webhooks]
Compressed Logs (/var/log/ltrack)
Runtime Config Hot Reload
SIEM Systems (Elastic/Splunk)
CLI Interface (Policy Management)
Metrics Endpoint (Prometheus Format)
[Command Controls]
[HTTP/HTTPS Exposure]
Audit Trail (Execution Logs)
Performance Dashboard

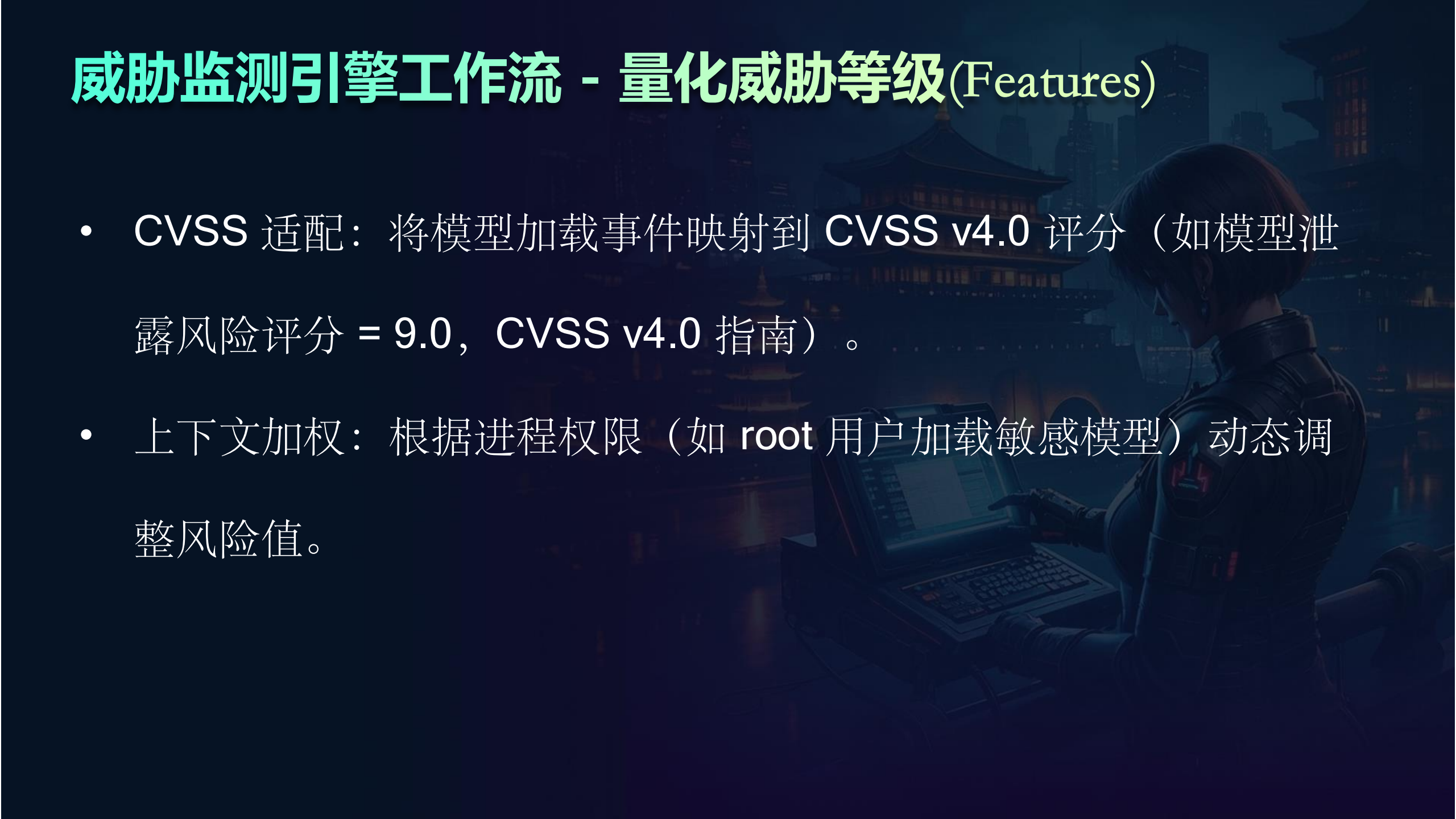
威胁监测引擎工作流 - 事件采集

- **文件操作**：挂钩security_file_open、security_mmap_file，捕获模型文件（.pt/.h5）的加载路径与哈希值。
- **进程间通信**：监控bpf_trace_printk跟踪进程树（如 Docker 容器内python进程加载模型）。
- **网络行为**：通过skb事件捕获sendfile传输的敏感文件（如模型权重外传）。

威胁监测引擎工作流 - 规则匹配

- 静态规则：基于 toml 配置文件约束监测范围。
- 依赖链分析：标记非标准依赖路径（如/tmp/libcustom.so）。

威胁监测引擎工作流 - 量化威胁等级(Features)



- CVSS 适配：将模型加载事件映射到 CVSS v4.0 评分（如模型泄露风险评分 = 9.0，CVSS v4.0 指南）。
- 上下文加权：根据进程权限（如 root 用户加载敏感模型）动态调整风险值。

威胁监测引擎工作流 - 动态阻断（实时防护）(Features)

- 进程终止：通过kill系统调用终止恶意进程（需 CAP_SYS_ADMIN 权限）。
- 文件隔离：将恶意模型文件移动到沙箱目录（如/var/quarantine）。
- 网络拦截：通过 eBPF TC（Traffic Control）丢弃外传敏感数据包。

零侵入监测

- 全局可见性
  - 单点监控
  - 内核级数据源
- 低资源损耗
  - 共享内核
  - 无应用干扰
- 快速部署
  - 一键启动
  - 动态扩展

企业级集成

- Prometheus 指标导出
  - 与 Grafana 集成生成威胁分布仪表盘
- Splunk 日志管道配置
  - 日志格式兼容 CEF（Common Event Format），支持 HTTP Event Collector (HEC)。
- Kubernetes Operator 部署
  - 与 Prometheus Operator 联动，自动生成 ServiceMonitor

轻量集成

组件	集成方式	关键能力	合规性支持
Prometheus	Pull模式（REST API）/Push模式（Pushgateway）	实时指标可视化、告警规则定义（需配合Alertmanager路由）	NIST SP 800-190
Splunk	支持Push/Pull模式（如CEF日志、API拉取）	威胁狩猎、数据分析与告警	GDPR Art.30
Kubernetes	Operator 自动化	零侵入式部署（如cAdvisor集成）、自动化服务发现	CIS K8s Benchmark

未来演进

- GPU检测支持
- 轻量集成支持
- 高度自定义事件类型

谢谢