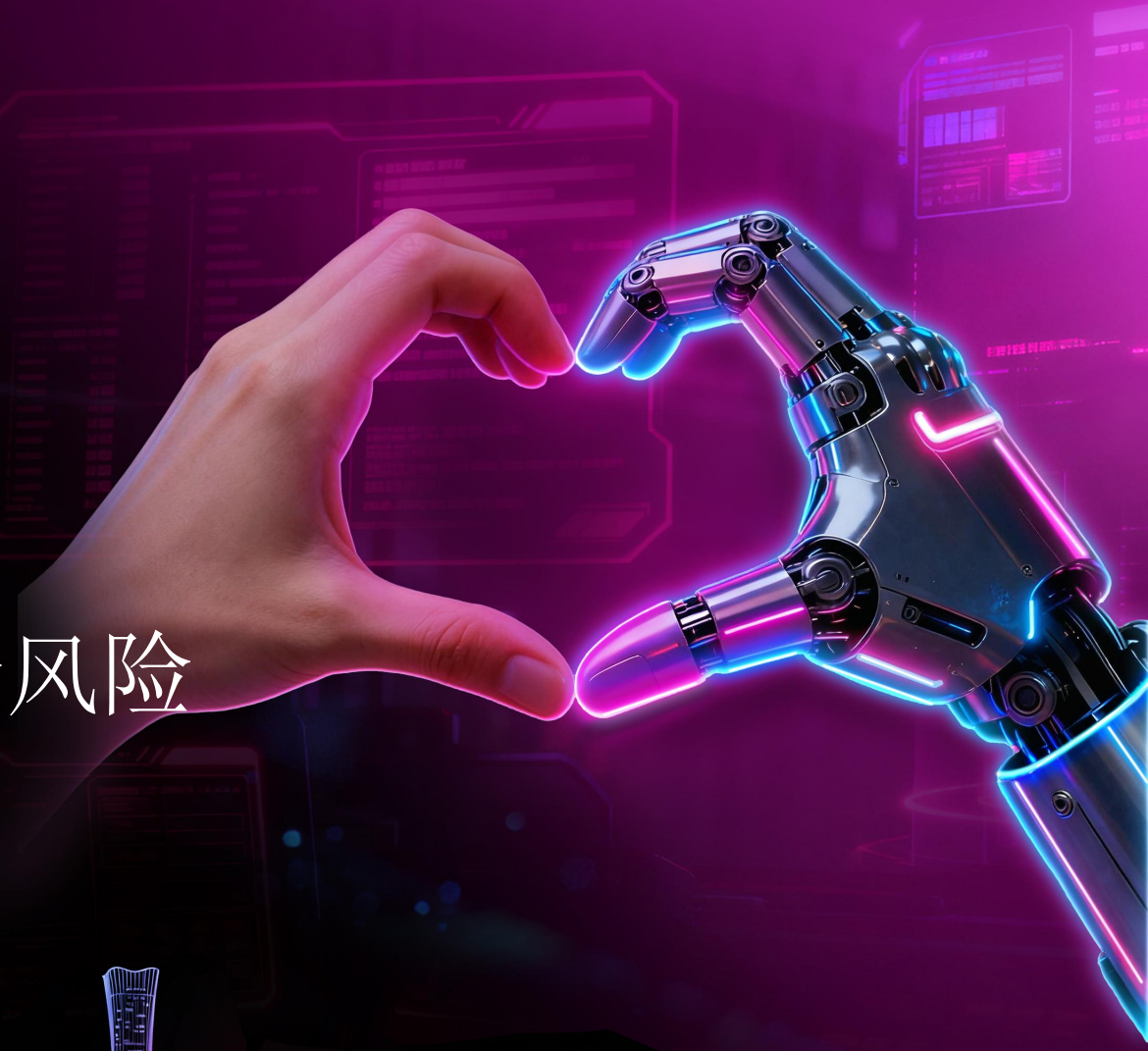
开源社
kaiyuanshe

COSCon'25
第十届中国开源年会

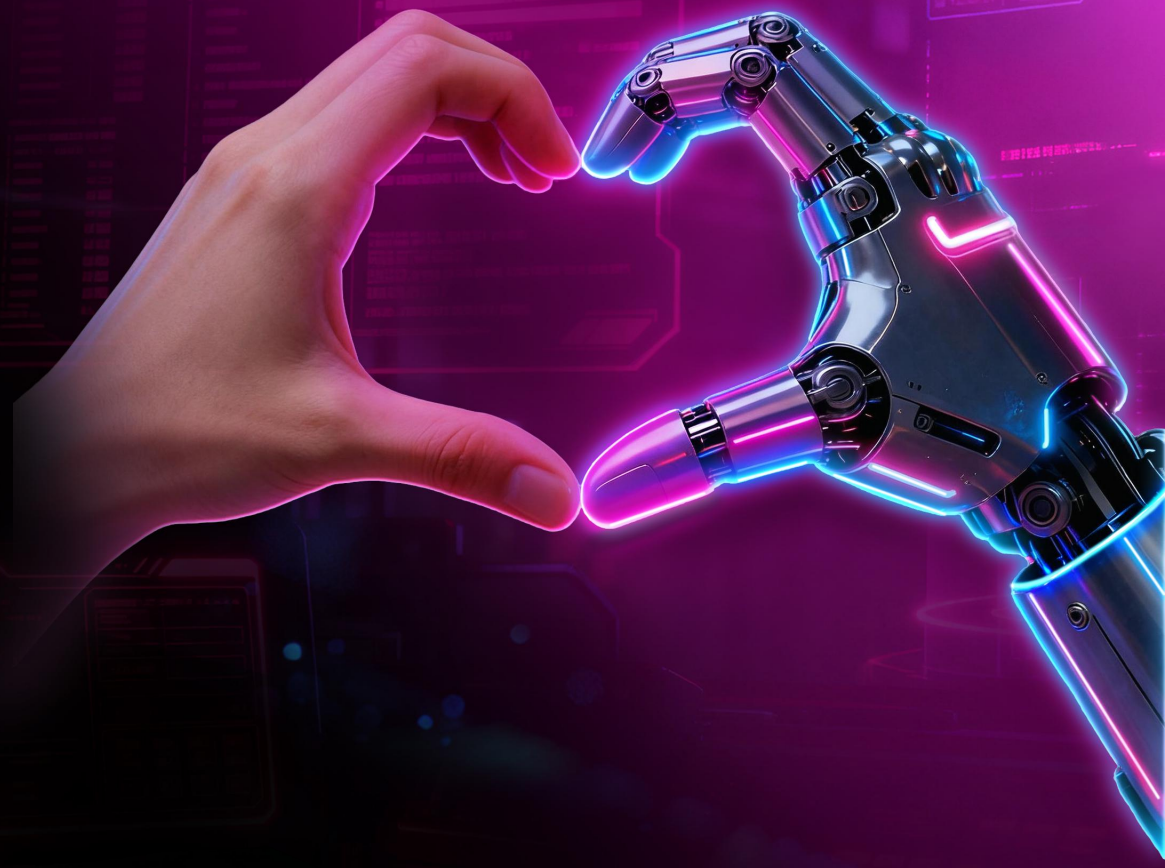
众智开源 | Open Source, Open Intelligence



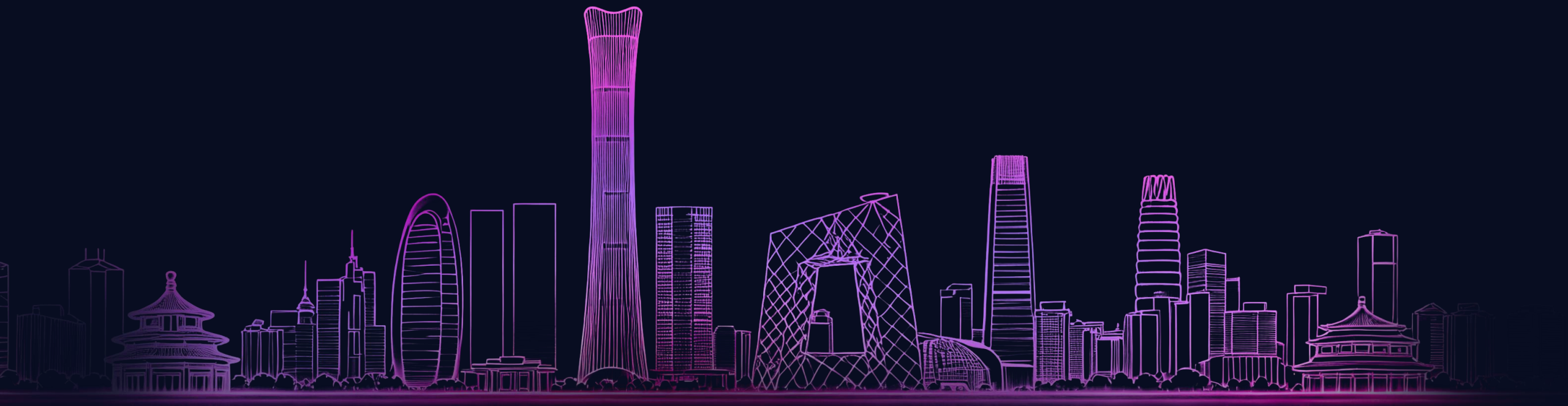
开源 AI大模型的合规与安全风险

王永雷 /Black Duck 高级软件安全架构师

CONTENTS



PART 01 开源大模型的现状



大模型的组成

开源社 kaiyuanshe

LLM

Models

+

Datasets

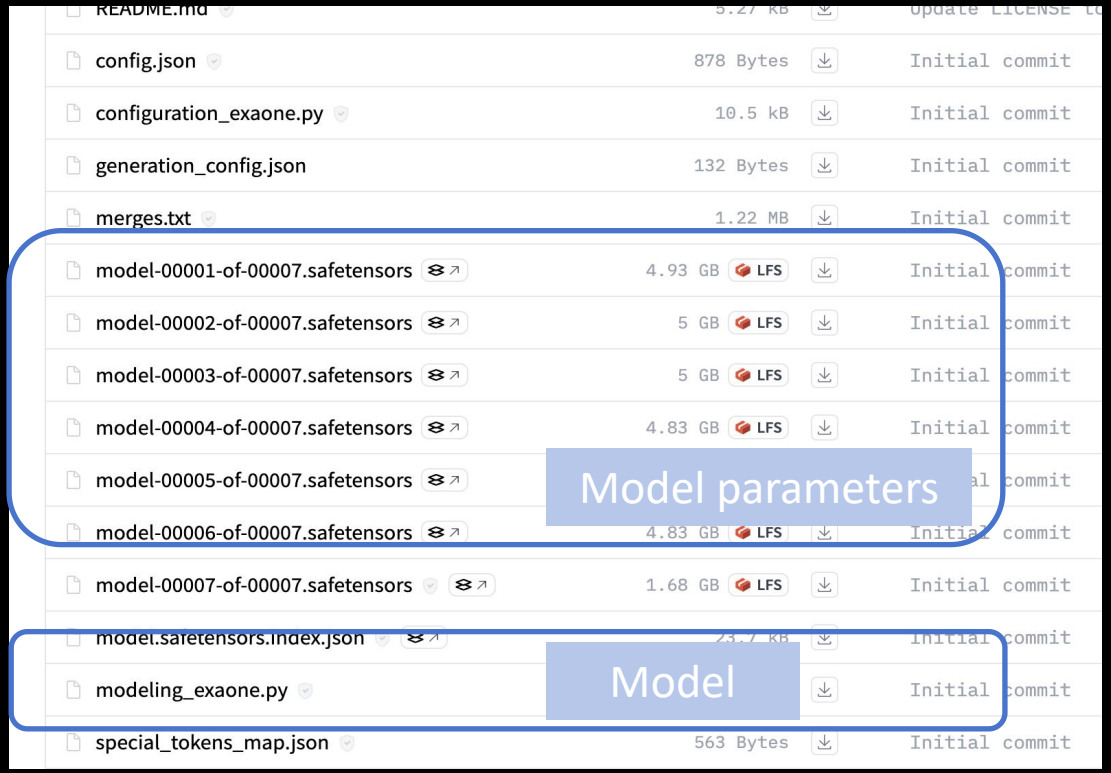
LLM Project

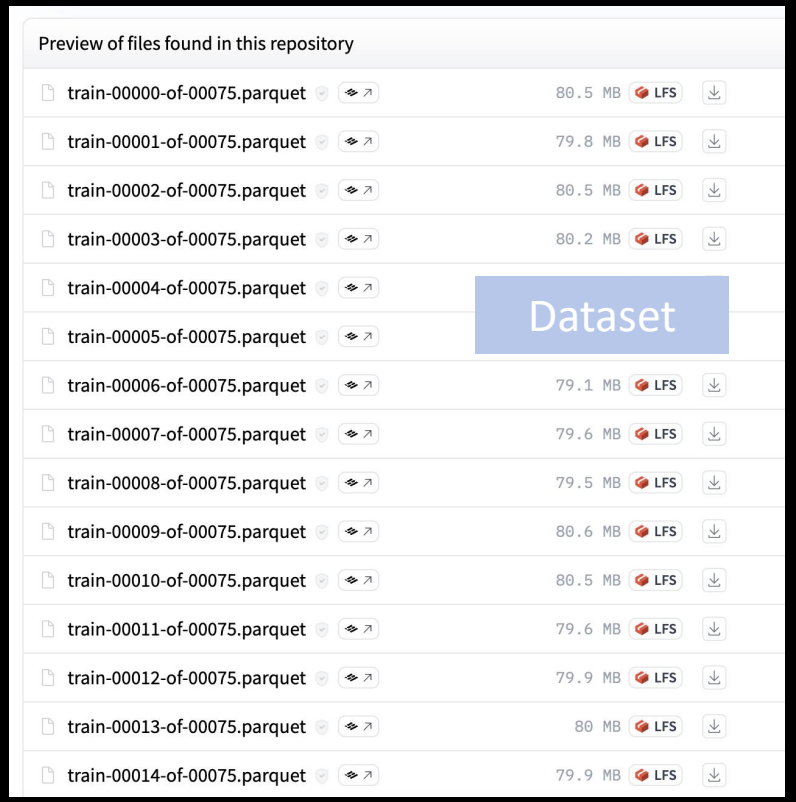
Model

Model parameters

+

Dataset





开源大模型的开源情况

开源社 kaiyuanshe

Model	Model/Weights	DataSets	Comments
Qwen3-Omni	Apache 2.0	无公开专属许可，训练数据含公开合规语料	2025 年登顶 Hugging Face 开源榜单的全模态模型，支持文本、图像、音频、视频处理，许可宽松可无限制商用，衍生模型超 9 万个
Qwen2.5 - MAX（阿里通义）	Apache 2.0	无公开专属许可，覆盖 119 种语言的公开数据源	中文语义理解优化突出，支持东南亚多语种，开源生态活跃，适合跨境电商等商用场景，许可无商用门槛
DeepSeek-R1	MIT	无公开专属许可，未披露完整数据集细节	开源模型性能标杆,MIT 许可下商用无额外门槛，仅未公开训练数据集全貌
DeepSeek 其他系列	自定义商业友好许可证	无公开专属许可，未披露完整数据集细节	除 R1 系列外，算法和权重许可禁止军事用途等场景；数据集未开放许可，且未公开训练数据的具体来源与授权细节。
LLaMA 4（Meta）	Llama 4 Community License	无公开专属许可，训练数据含 200 种语言的公开语料	采用混合专家架构，包含 Maverick 和 Scout 两个版本，许可为 Meta 自定义，商用需遵守社区许可条款，对学术研究友好
StableLM Alpha	CC BY - NC - SA 4.0	CC BY - NC - SA 4.0（仅非商用）	算法、权重与数据集均受非商业许可限制；数据集是基于“堆”扩展的 1.5 万亿 token 实验性数据集，微调模型仅限研究使用，商用需替换无限制数据集。

开源大模型的安全事件-供应链攻击：开源生态的〞特洛伊木马〞

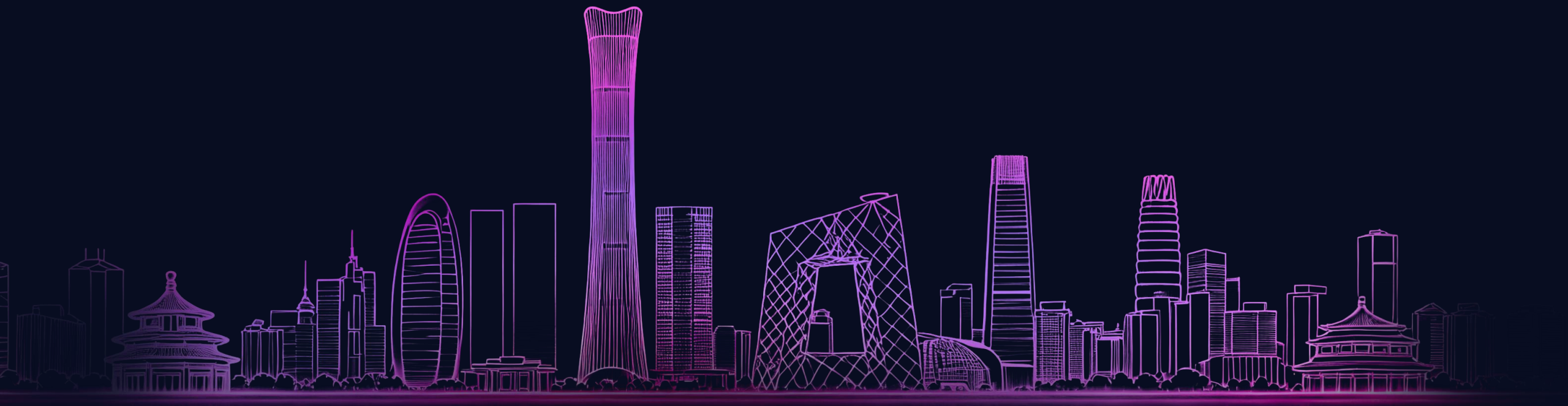
攻击者通过篡改开源代码库、模型仓库或依赖项，将恶意代码植入软件供应链。

攻击途径：

- PyPI/conda 等包管理器上传恶意版本
- Hugging Face 等模型平台植入后门模型
- 修改开源代码库后提交恶意 PR
- 劫持模型命名空间，在可信名称下部署恶意模型

Date	Incidents	Comments
2023/03	Meta LLaMA 泄露	研究团队模型意外公开，引发安全风险，后被用于开发多款闭源商业模型
2024 /12	YOLOv11 投毒事件	v8.3.41 和 v8.3.42 版本被植入加密挖矿软件，下载量超 96 万，用户设备被感染
2025 / 2	Hugging Face 平台 "nullifAI" 攻击	攻击者利用 "损坏的 pickle 文件" 技术绕过安全检测，上传含恶意代码的模型，可收集用户设备信息
2025 / 8	Nx 构建工具被黑	植入恶意代码，主动调用本地 AI 工具窃取信息，攻击数百万开发者

PART 02 欧盟AI Act和 CRA法规



欧盟 AI Act 开源大模型监管要点与合规边界

https://artificialintelligenceact.eu/high-level-summary/

开源社 kaiyuanshe

监管的要求

通用人工智能模型（GPAI）

通用人工智能模型（GPAI）具备广泛应用能力，是能够在多种任务中表现出色的开源基础模型。如LLaMA 3凭借其在自然语言处理方面的强大能力，可应用于文本生成、智能问答等多个领域；Qwen 2.5也在语言理解和生成上有着卓越表现，能为用户提供高质量的语言交互服务 。

系统性风险通用AI模型（GPAISR）

- 训练累计计算量：当模型训练累计计算量 ≥ 10^{25} FLOPS时，其强大的计算能力和广泛的数据处理范围可能引发不可控的风险，因此被视为GPAISR的判定标准之一。
- 顶尖能力与关键领域影响：若模型经评估具备“顶尖能力”，并对医疗、金融、司法等关键领域产生影响，因其决策可能直接关系到人们的生命健康、财产安全和社会公平正义，所以也符合GPAISR判定标准。
- 欧盟AI办公室综合认定：欧盟AI办公室依据专业的评估体系和丰富的经验，对模型进行全面审查和综合考量，若认定某模型具有系统性风险，则将其归为GPAISR。

欧盟 AI ACT 法案开源大模型透明度要求细则

算法·权重·数据集报备要求全解析

开源社 kaiyuanshe

要求类型	普通开源 GPAI 模型	开源 GPAISR 模型	Comments
算法 (架构) 报备	☑豁免 (已公开即可)	❌不豁免 (必须报备完整说明)	
权重参数报备	☑豁免 (已公开即可)	❌不豁免 (必须公开全部权重)	
数据集完整报备	❌不要求 (仅需摘要)	❌不要求 (仅需摘要)	• 数据来源与收集方式 • 版权内容占比 • 数据多样性指标 • 敏感数据使用情况
数据集摘要报备	☑必须 (使用官方模板)	☑必须 (更详细，需向 AI 办公室报备)	
模型文档表填写	☑豁免	❌不豁免 (必须填写并保存 10 年)	

EU CRA · 网络弹性法案概览 //数字产品安全基线

- 软件：如ISV
- 硬件：如IoT设备、智能手机、计算机
- 服务：如基于云平台远程控制智能家居设备

所有“带有数字元素的产品”（PDE）
包括任何软件或硬件及其远程数据处理方案

终端用户软件	软件即服务(SaaS)	硬件运行支撑服务 (如语音识别引擎)	嵌入式软件	网络访问型软件 (如API接口)	AI模型

- 几乎所有具备联网等数字化功能的电子类产品，包括电视、冰箱、智能音响等
- IoT是典型的适用领域
- AI模型

EU CRA · 关键要求 · 软件安全部分

开源社 kaiyuanshe

安全设计

- 产品必须在安全的开发生命周期内开发 **//S-SDLC**
- 要确保对存储和传输中的数据进行强加密
- 实施安全启动过程以在执行前验证软件完整性
- 设计产品时要**尽量减少攻击面**并防止篡改

漏洞处理与事件报告

- 制造商必须有**管理漏洞流程**，包括：定期更新和打补丁；检测到“任何被大肆利用的漏洞”后**24h内**向欧盟网络安全局（ENISA）披露

SBOM与技术文档

- 制造商必须准备并维护全面的技术文档以证明合规
- 需维护最新的软件物料清单（**SBOM**），列出代码库中所有开源代码组件和依赖、第三方和自有组件及依赖项等信息

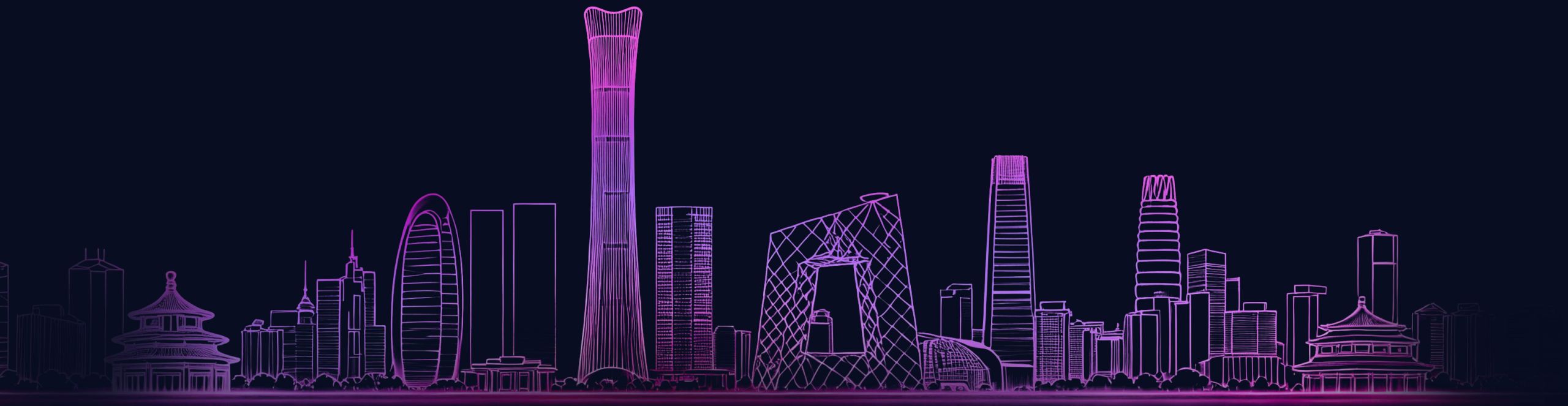
更新与维护

- 制造商必须确保在产品预期寿命内提供安全更新并给予支持

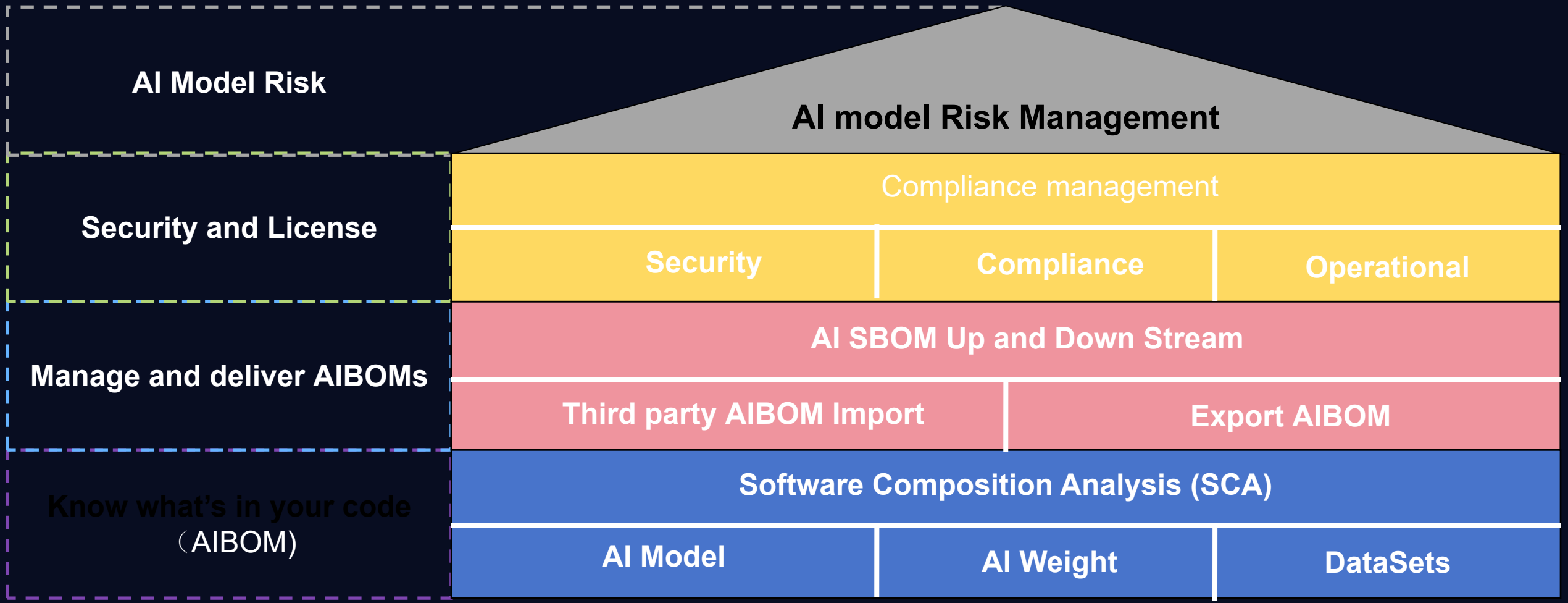
用户说明

- 必须向用户提供清晰易懂的信息，说明产品使用中网络安全的相关风险
- 产品必须有标签，告知用户有关网络安全功能及其更新情况

PART 03 如何应对挑战思考



LLM-Risk Management



AI ACT合规行动清单

开源社 kaiyuanshe

Model	Action List	Comments
普通开源 GPAI 模型	1：验证开源许可类型（确保满足豁免要件） 2：公开模型架构、权重、使用说明 步骤 3：按欧盟模板编制并公开训练数据摘要 步骤 4：留存合规证明文件（许可文本、公开记录）	• 核心原则：差异化合规，不搞“一刀切”普通开源模型：公开 = 合规，仅需补充数据摘要 • 开源 GPAISR：全流程报备，无豁免空间 • 关键边界：开源≠免合规，数据摘要公开是“硬要求” • 实操优先级：先判定风险等级→再验证豁免条件→最后落实对应要求 • 长期关注：欧盟动态调整（如风险阈值、模板内容更新）
开源 GPAISR 模型	1：提前向欧盟 AI 办公室报备技术文档（算法 + 权重 + 数据摘要） 2：签署《通用人工智能行为准则》（准强制要求） 3：建立风险监测与事件报告机制（24 小时重大事件上报） 4：接受第三方独立评估	

供应链合规管理

- 生成标准化 SBOM 并开展第三方依赖尽职调查。
- 建立组件更新淘汰机制

SBOM标准化

通过标准化生成软件物料清单，统一格式与结构，确保组件信息可追溯。

组件全记录

完整记录开源组件及其版本，涵盖所有依赖项，提升透明度。

依赖尽职查

对第三方组件进行安全与许可审查，识别潜在法律与漏洞风险。

风险及时识

自动检测高危组件与过期依赖，快速定位问题源头。

动态生命周期

建立组件全生命周期管理机制，跟踪各阶段使用状态。

高危更新替

推动高风险依赖的升级或替换，降低安全暴露面。

淘汰机制强

制定过时组件下线流程，防止陈旧组件继续使用。

供应链清洁

持续维护软件供应链安全，保障系统长期可控与合规。

EU CRA合规行动清单-漏洞管理体系的搭建

开源社 kaiyuanshe

漏洞管理体系的搭建

- 制定透明漏洞披露政策和修复机制。
- 接入开源漏洞共享平台同步威胁情报

披露政策制定

建立公开透明的漏洞披露流程，明确报告渠道与响应时限。

快速修复机制

设定漏洞分级响应标准，确保高危漏洞在规定时间内修复。

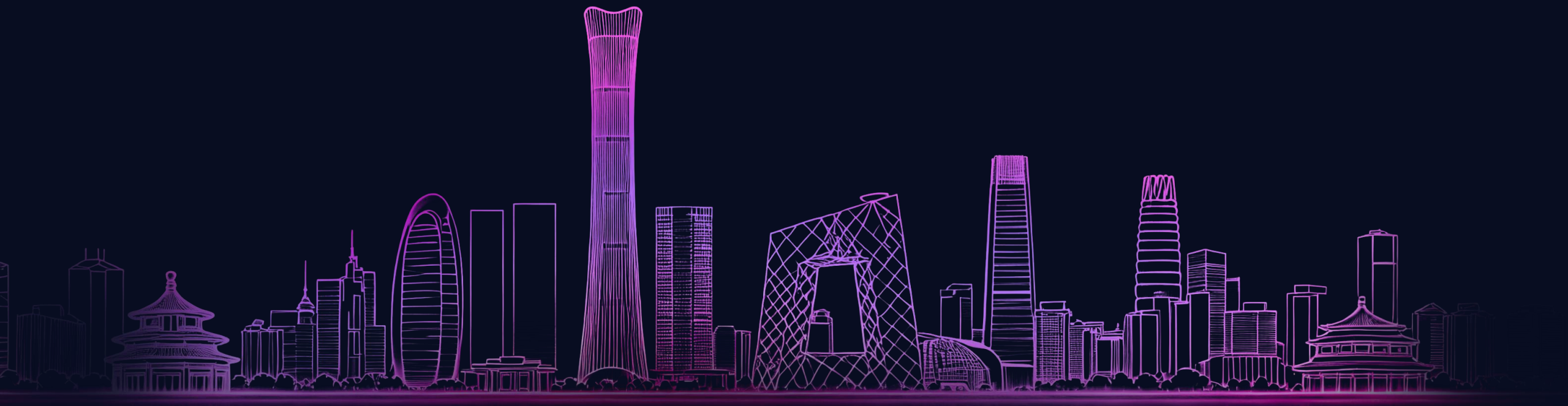
威胁情报同步

接入主流开源漏洞库，实时获取并共享最新安全威胁信息。

协同响应网络

与社区及第三方平台联动，提升漏洞处置效率与覆盖范围。

PART 04 Q&A



开源社
kaiyuanshe

COSCon'25
第十届中国开源年会

众智开源 | Open Source, Open Intelligence

Thanks

永雷

Pudong New Area, Shanghai

王永雷