THALES
Building a future we can all trust

Sentinel
AI 模型保护

www.thalesgroup.com

应用案例 —— Lunit AI模型保护

医疗科技工具聚焦于使用AI增强癌症诊断 & 陪护.

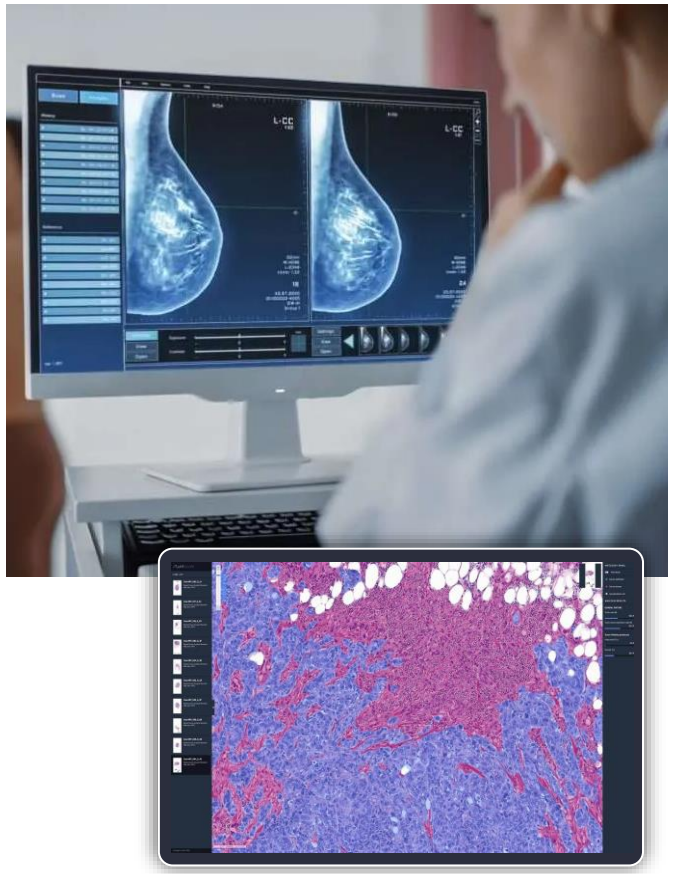
> **Lunit需要为其AI赋能的产品寻找软件授权及保护解决方案**

> **避免AI模型被盗用是当务之急**

> **若AI产品出现不准确的输出，其影响甚大**

- 误诊 & 医疗事故
- 合规违规
- 声誉影响

Lunit



> 解决方案

- 通过使用Sentinel Envelope外壳保护Lunit的AI应用栈
- 保护AI模型以避免对模型输入/输出攻击
- 通过Sentinel SDK 完成对AI赋能的功能进行访问控制

Lunit使用Sentinel软件货币化方案实现了业务增长，为将来实现盈利及IPO成功铺路

为什么AI保护如此重要?

> 声誉损失

- AI模型存在被恶意滥用的问题，这可能导致误诊、越权、不恰当的输出结果及品控投诉等，最终造成对AI及开发商的名誉损失。

> 合规罚款

- 在合规约束的相关行业，对AI的滥用将导致合规处罚，例如医疗设备行业是一个高度合规监管的行业，除了对患者的医疗安全有所考虑，合规要求也体现在了对数据隐私保护、职工安全和管理责任上。

> IP 盗窃

- 因为开发AI模型是一个消耗大量时间、资源的高成本项目，对AI模型进行保护将有助于开发商保护其投资,并保持竞争优势。

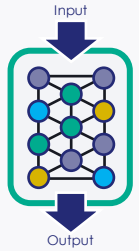
在不受控环境下部署AI产品将导致威胁增加

核心

边端

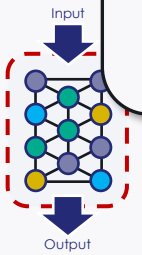
公有云

- 公有云数据中心是更为安全的环境
- AI滥用发生的可能性**较低**



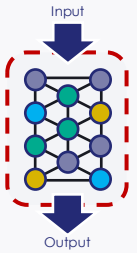
私有云

- 不受控的数据中心，安全性未知
- 无法保证访问权限
- AI滥用发生的可能性**高**



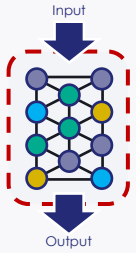
本地化部署

- 物理上可及
- 没有或欠缺对生产环境的控制
- AI滥用发生的可能性**高**



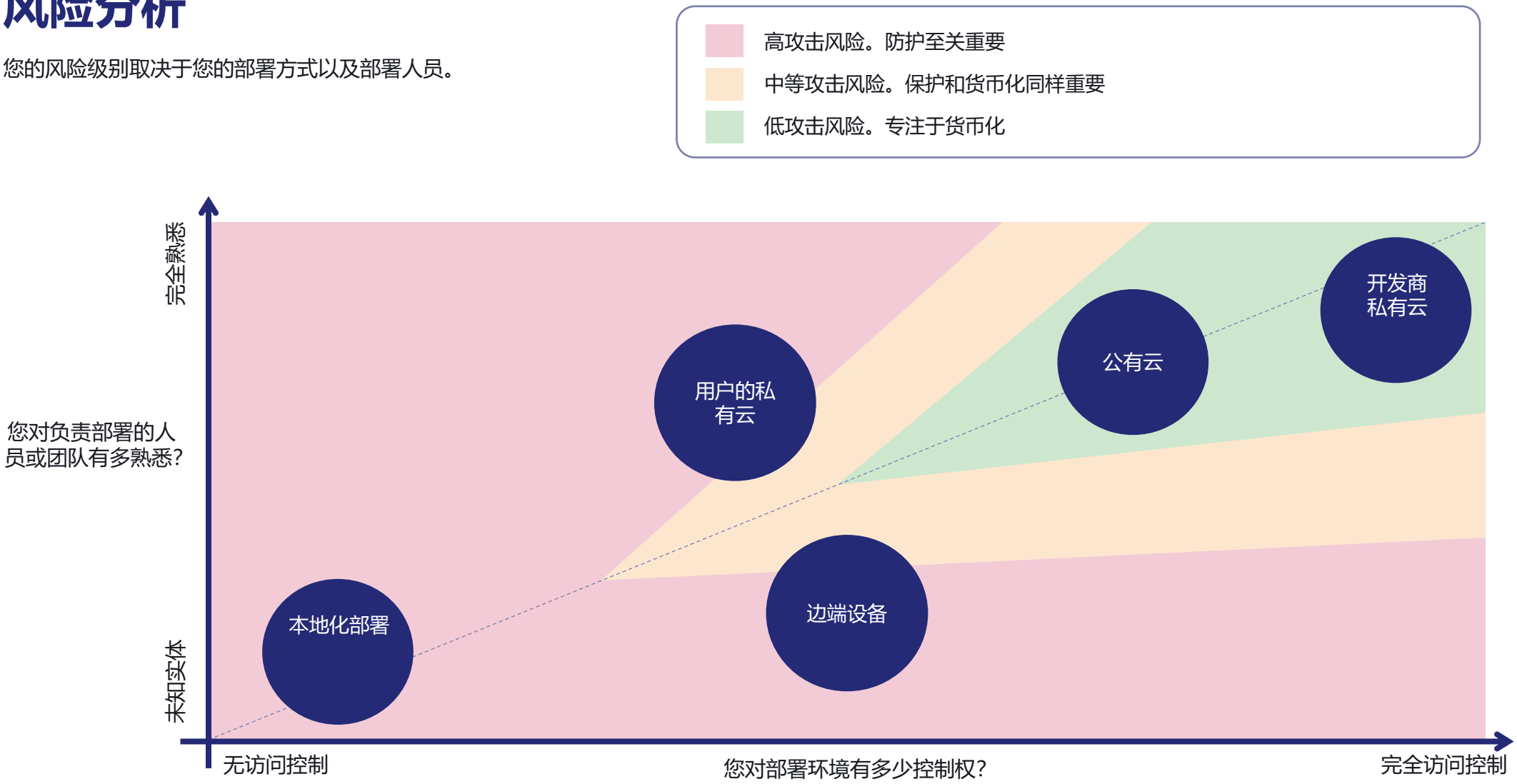
边端设备

- 物理上可及
- **难以管理和监控**
- AI滥用发生的可能性**高**



风险分析

您的风险级别取决于您的部署方式以及部署人员。



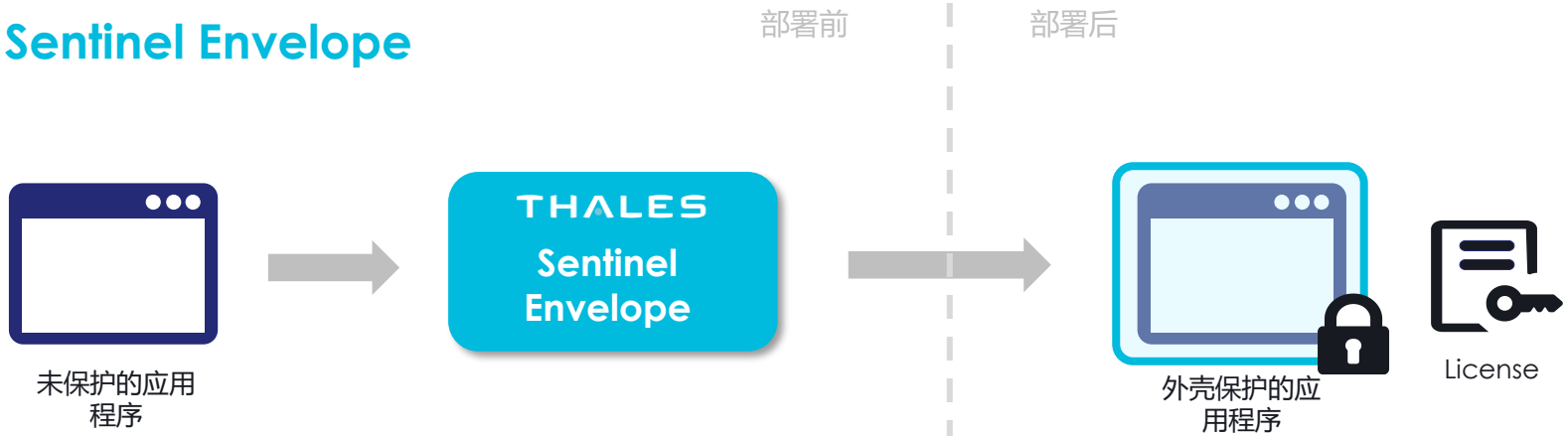
部署到不受控制的环境中时针对AI的威胁

	应用程序篡改	模型篡改	模型盗窃
这是什么?	• 对支持AI的应用程序进行逆向工程，以操纵应用程序与模型之间的输入和输出。 • 为创建软件的非法副本打开大门	• 操纵人工智能模型以引入漏洞、偏见或错误	• 提取人工智能模型并在不同的解决方案中重复使用它们
风险	• 人工智能模型输出错误 • 应用程序知识产权被盗，收入损失	• AI模型输出错误	• 让竞争对手轻松获得优势
场景举例	• 错误分类图像，绕过安全系统 • 绕过验证检查并批准未经授权的交易 • 以任何形式操纵输出	• 错误分类图像，绕过安全系统 • 虚假医疗诊断 • 对IT攻击的不当反应	• 供应商窃取竞争对手的 AI 模型用于自己的解决方案，从而允许他们以低得多的成本发布类似产品。

OWASP® 这些威胁在最新的 OWASP 十大机器学习安全风险中有描述

Sentinel AI保护组件介绍

> Sentinel Envelope



- 一旦完成保护，应用程序及模型文件将只能在有效license存在的前提下运行
- 信息包括了可以被允许的功能，以及任何追加的license策略，如到期时间和受允许的用户数量等
- 不同的应用程序和模型可套用其自身所需的不同的license，或相同的通用license

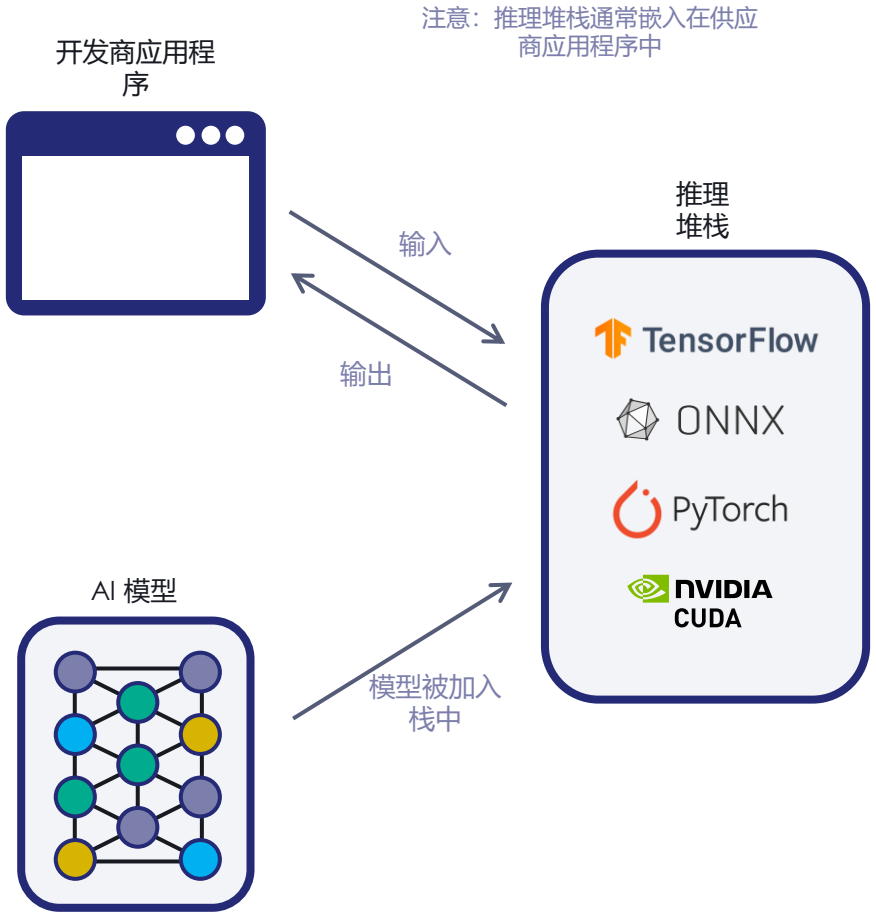
> Sentinel 数据文件保护模块

THALES
Building a future we can all trust

AI 攻击的场景&
Sentinel 解决方案

www.thalesgroup.com

AI 组件说明



> 开发商应用程序

- ISV 应用程序利用 AI 作为其产品的一部分，通常用于增强或扩展功能
- 例如，医疗应用利用人工智能分析扫描仪图像来提高诊断率

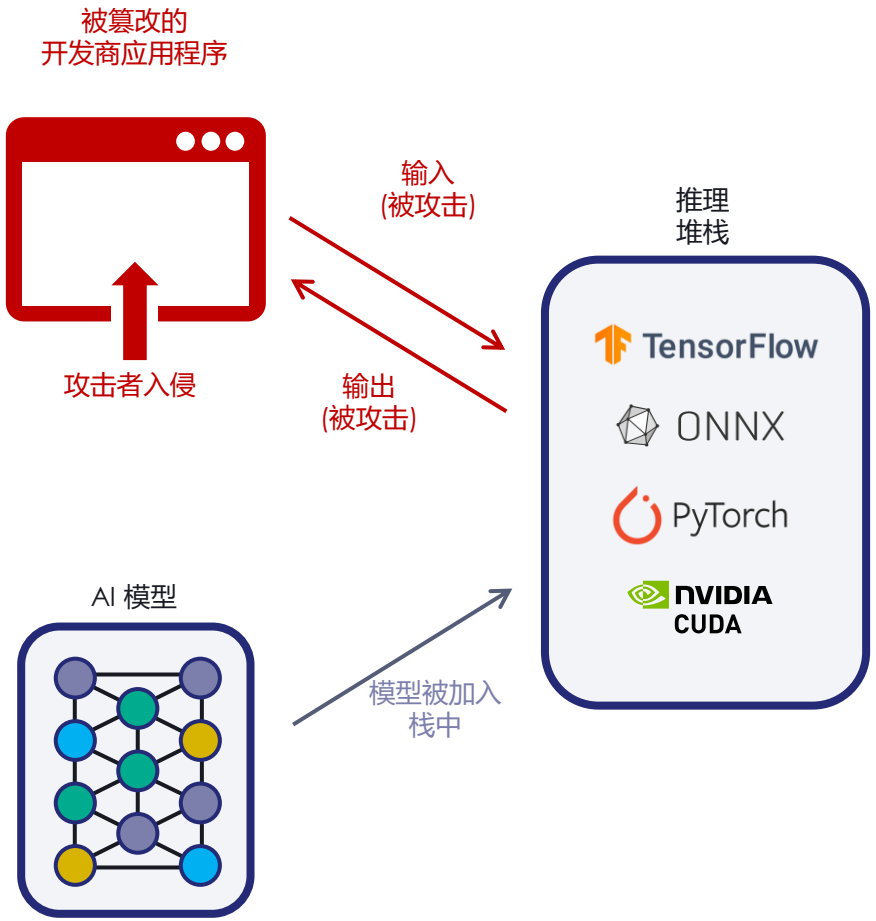
> 推理堆栈

- 推理引擎加载、解释和执行模型（例如生成预测或分类）
- 运行时管理开发商应用程序和 AI 模型之间的数据流，实时处理输入、处理和输出

> AI 模型

- 描述模型及其所有参数和结构的文件。

攻击方式 1 –开发商应用程序篡改

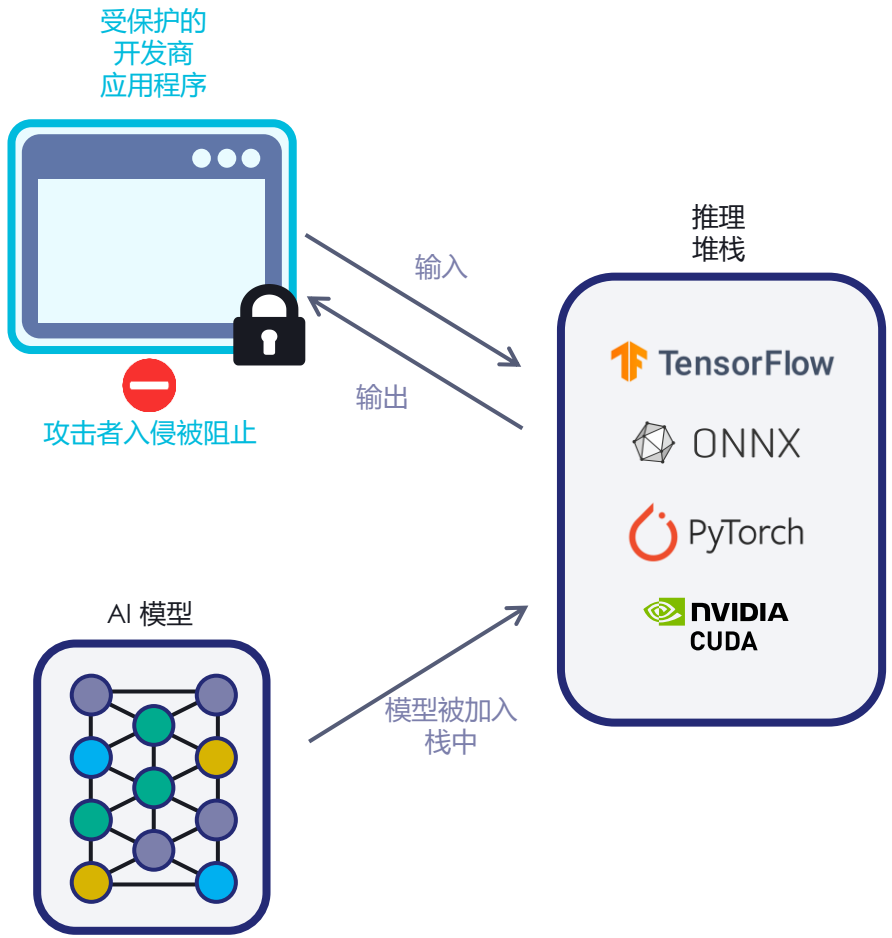


> 开发商应用程序被篡改，损害模型的输入和输出

> 此类威胁将导致：

- 输入操纵/对抗攻击
  - 攻击者发送被操纵的输入数据，利用模型解释中的漏洞，误导模型做出错误的预测或分类
- 后门注入
  - 利用带有注入信息的输入在训练期间将隐藏的触发器植入模型
- 输出完整性攻击
  - 强制应用程序在不改变输入数据的情况下扭曲结果

攻击方式 1 解决方案 – 保护开发商应用程序以阻止逆向工程



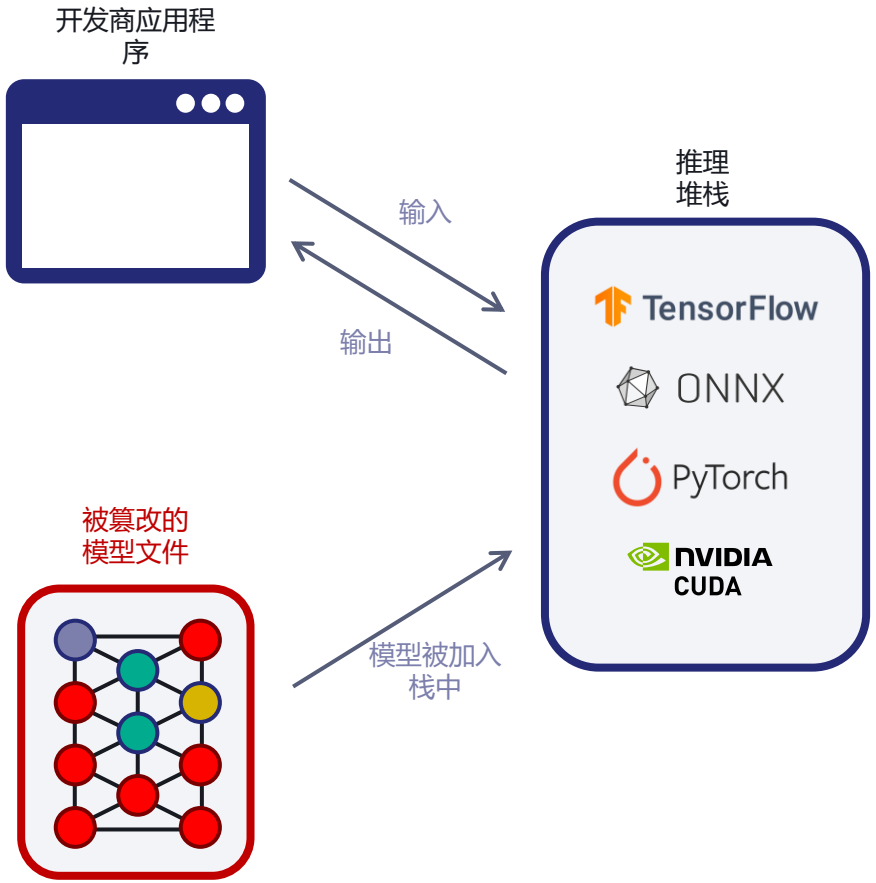
> Sentinel Envelope 被用于保护开发商应用程序

- 注意:根据环境/设置，推理堆栈也可能包含在外壳内

> 好处

- 保护应用程序代码免遭反汇编和逆向工程
- 保护验证输入的代码，防止输入逻辑被修改。
- 防止应用程序端输出完整性攻击，因为处理输出的应用程序代码受到保护。
- 允许在应用程序内部安全地实现单独的保护代码，以防止模型反转和模型提取攻击

攻击方式 2 – 模型文件篡改

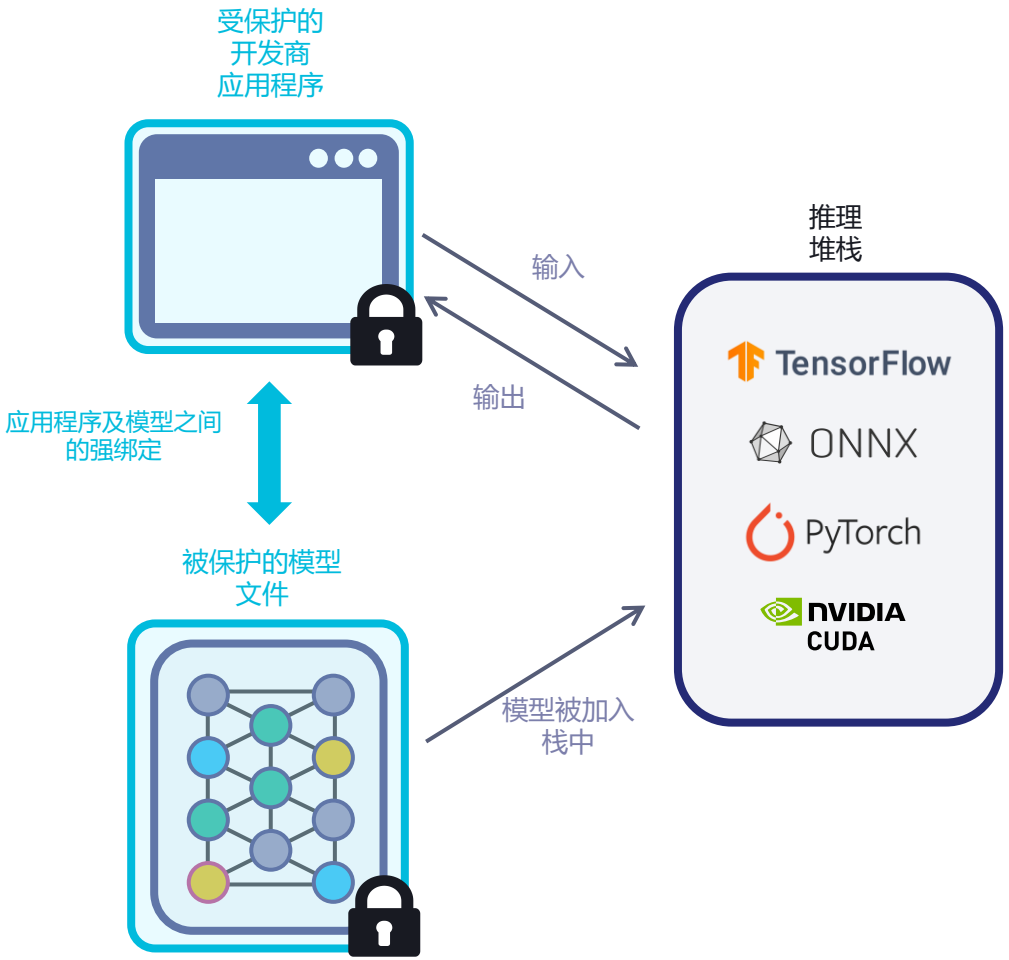


> 篡改模型参数信息将导致不可预见的行为

> 此类威胁将导致：

- 模型中毒
  - 操纵模型参数和权重，产生非预期结果和响应
- 迁移学习攻击
  - 攻击者使用恶意数据集重新训练模型，使其行为发生改变
- 模型反转攻击
  - 对已训练的模型数据进行逆向工程，以提取敏感信息

攻击方式 2 解决方案 – 模型文件加密保护

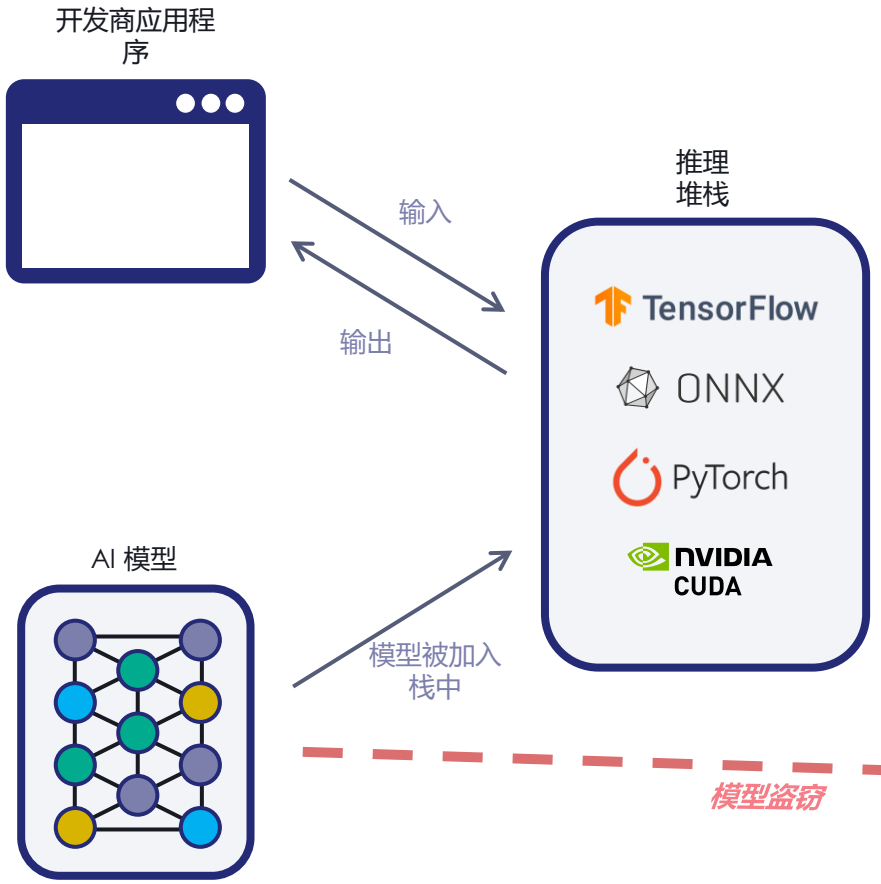


> Sentinel 数据文件保护模块用于模型文件的保护

> 好处

- 保护模型免受中毒，因为不再可能存在针对性的参数/权重/偏差更改。
- 如果无法访问模型参数/权重/偏差，迁移学习攻击就无法进行。
- 由于无法访问模型，模型反转和模型提取攻击得以阻止，迫使开发商应用程序成为进入模型的唯一途径。

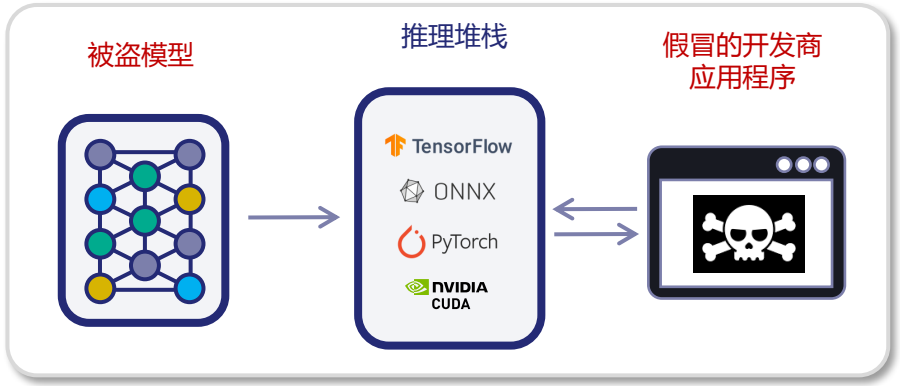
攻击方式 3 – 模型盗窃



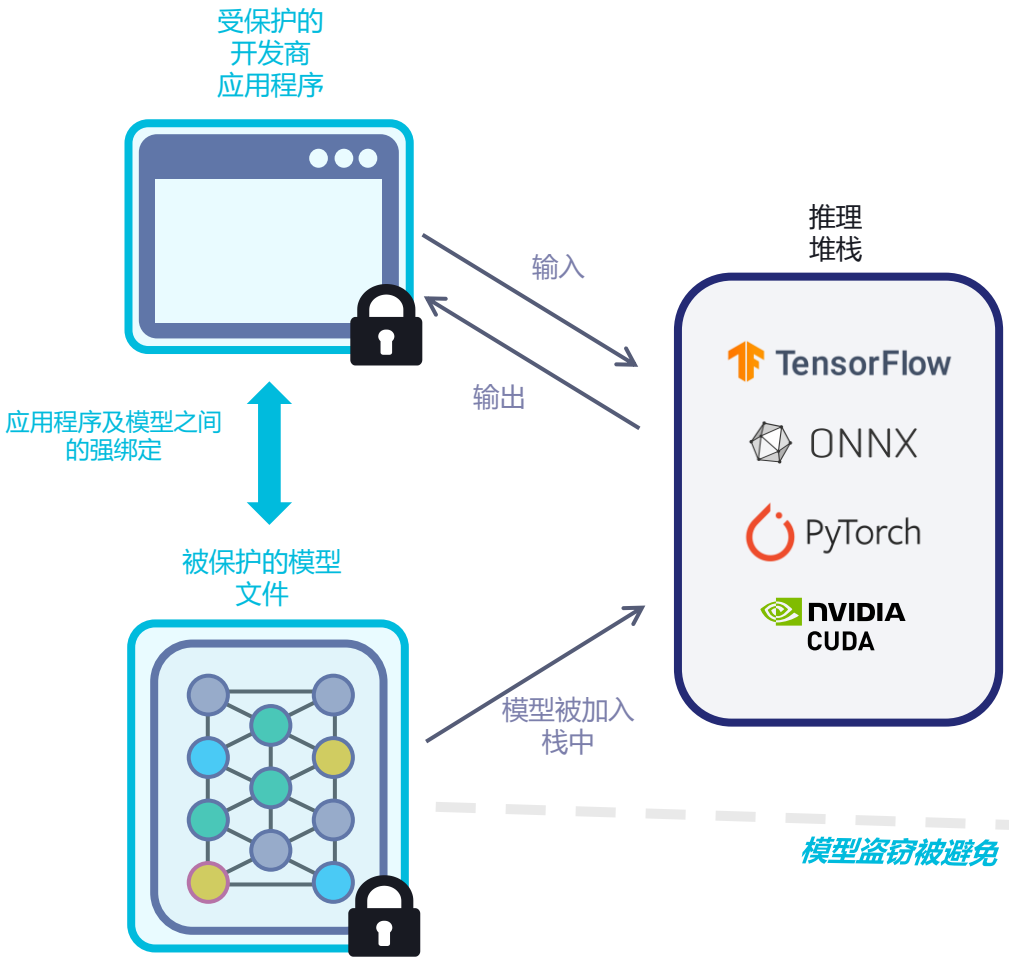
> 模型被恶意提取、复制和重复使用

> 此类威胁将导致：

- 窃取模型权重/参数——通常是模型中最有价值的部分
  - 窃取模型算法
- 被盗模型可用于：
  - 以更低的成本和上市时间打造竞争产品
  - 提取敏感或私人数据



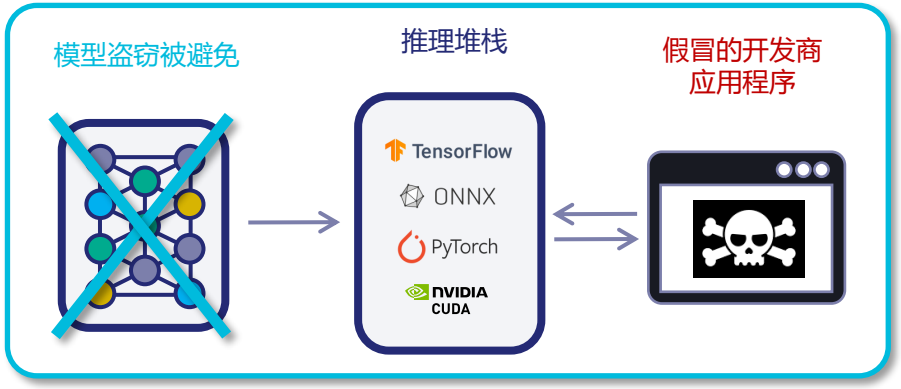
攻击方式 3 解决方案 – 保护模型文件并绑定应用程序



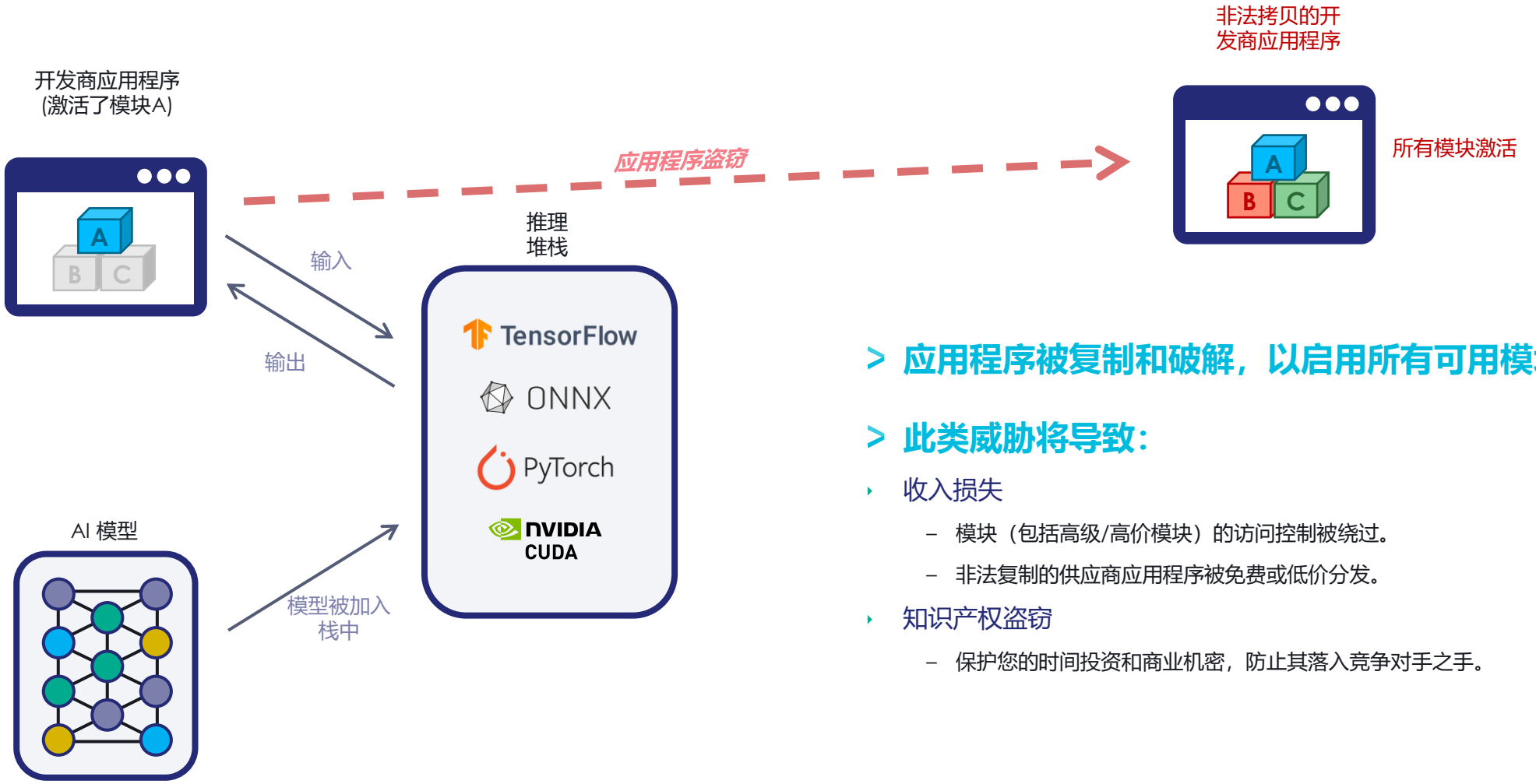
> Sentinel 数据文件保护模块对模型文件施加保护

> 好处

- 保护参数和权重中的机密信息不被提取
- 将模型绑定到应用程序，防止在未经授权的应用程序中加载。



攻击方式 4 – 软件盗版

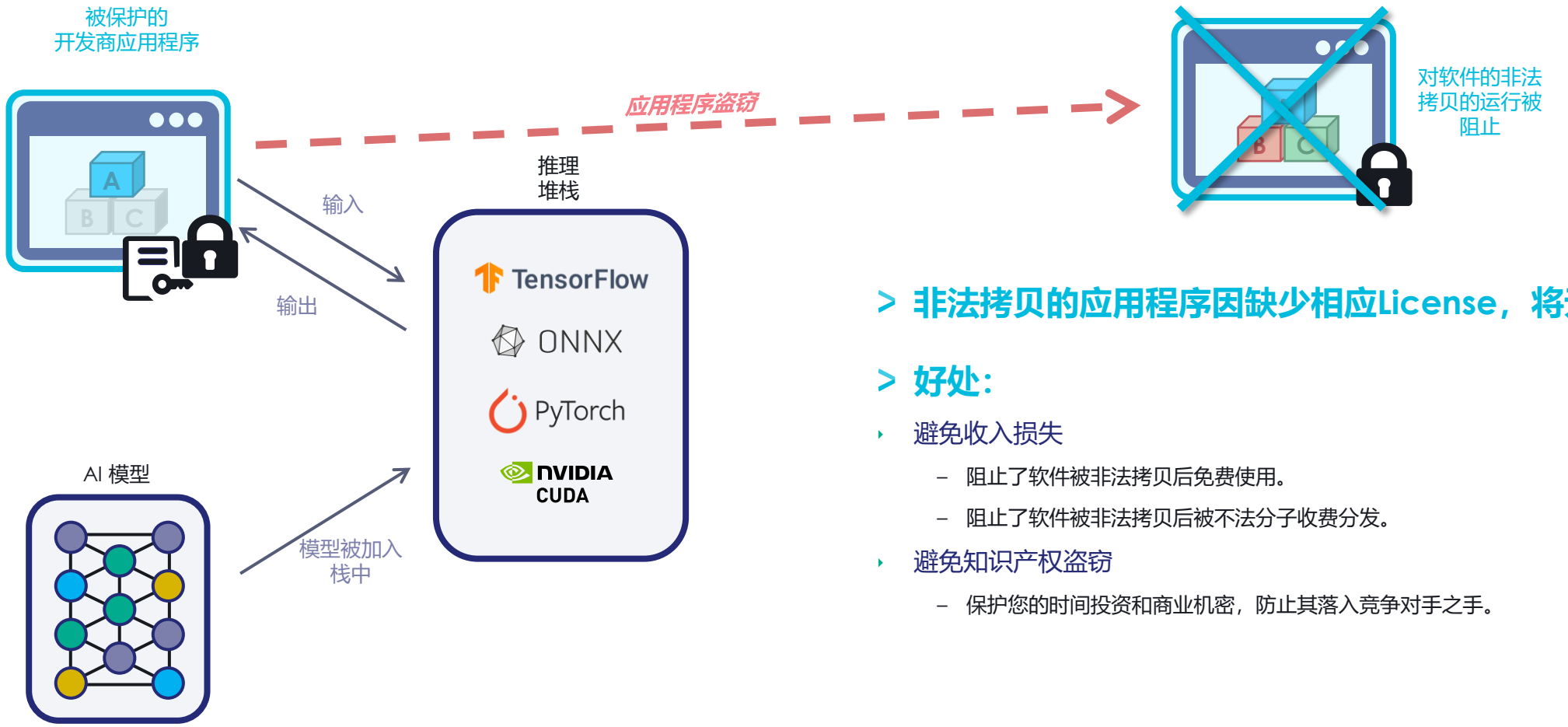


> **应用程序被复制和破解，以启用所有可用模块**

> **此类威胁将导致：**

- 收入损失
  - 模块（包括高级/高价模块）的访问控制被绕过。
  - 非法复制的供应商应用程序被免费或低价分发。
- 知识产权盗窃
  - 保护您的时间投资和商业机密，防止其落入竞争对手之手。

攻击方式 4 解决方案 – 保护应用程序并绑定license策略

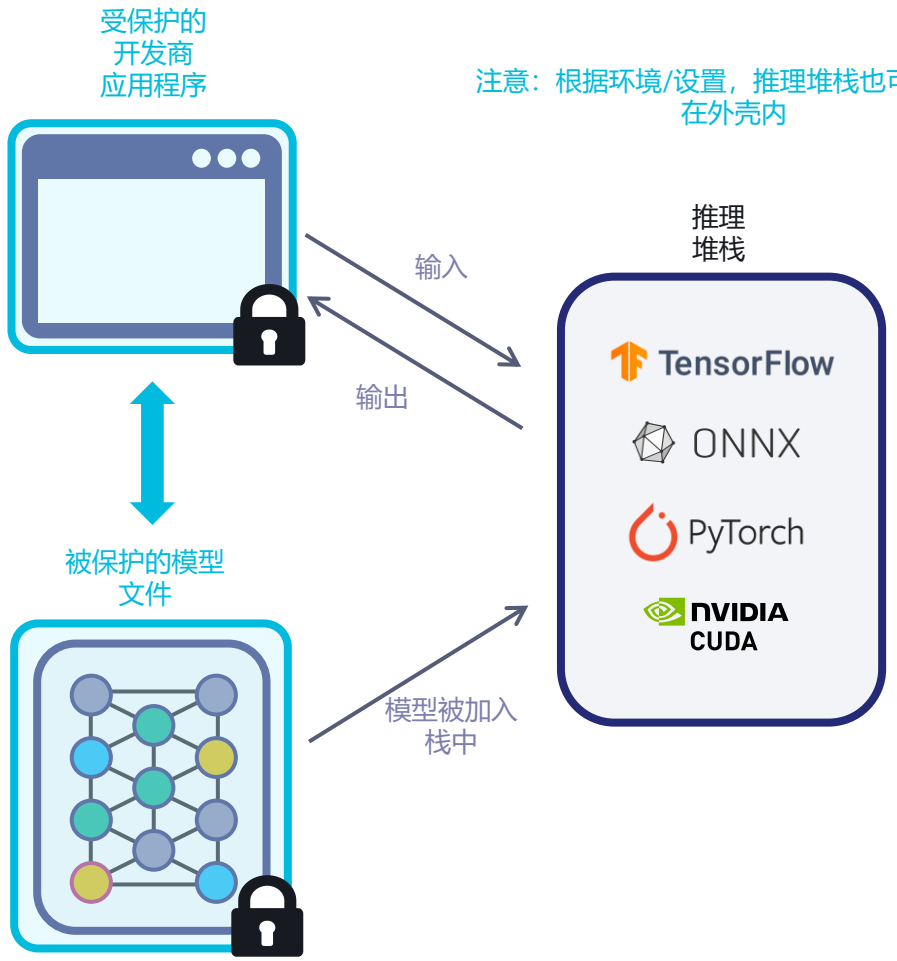


> 非法拷贝的应用程序因缺少相应License，将无法运行

> 好处：

- 避免收入损失
  - 阻止了软件被非法拷贝后免费使用。
  - 阻止了软件被非法拷贝后被不法分子收费分发。
- 避免知识产权盗窃
  - 保护您的时间投资和商业机密，防止其落入竞争对手之手。

Sentinel 解决方案总结 – 对开发商应用程序及AI模型文件的双重保护



使用Sentinel Envelope保护开发商应用程序

使用Sentinel 数据文件保护 加密模型文件

> 受保护的应用程序

- 保护应用程序代码免遭反汇编和逆向工程
- 防止盗版

> 加密模型文件

- 通过阻止针对性地更改参数/权重/偏差，保护模型免受不利修改。