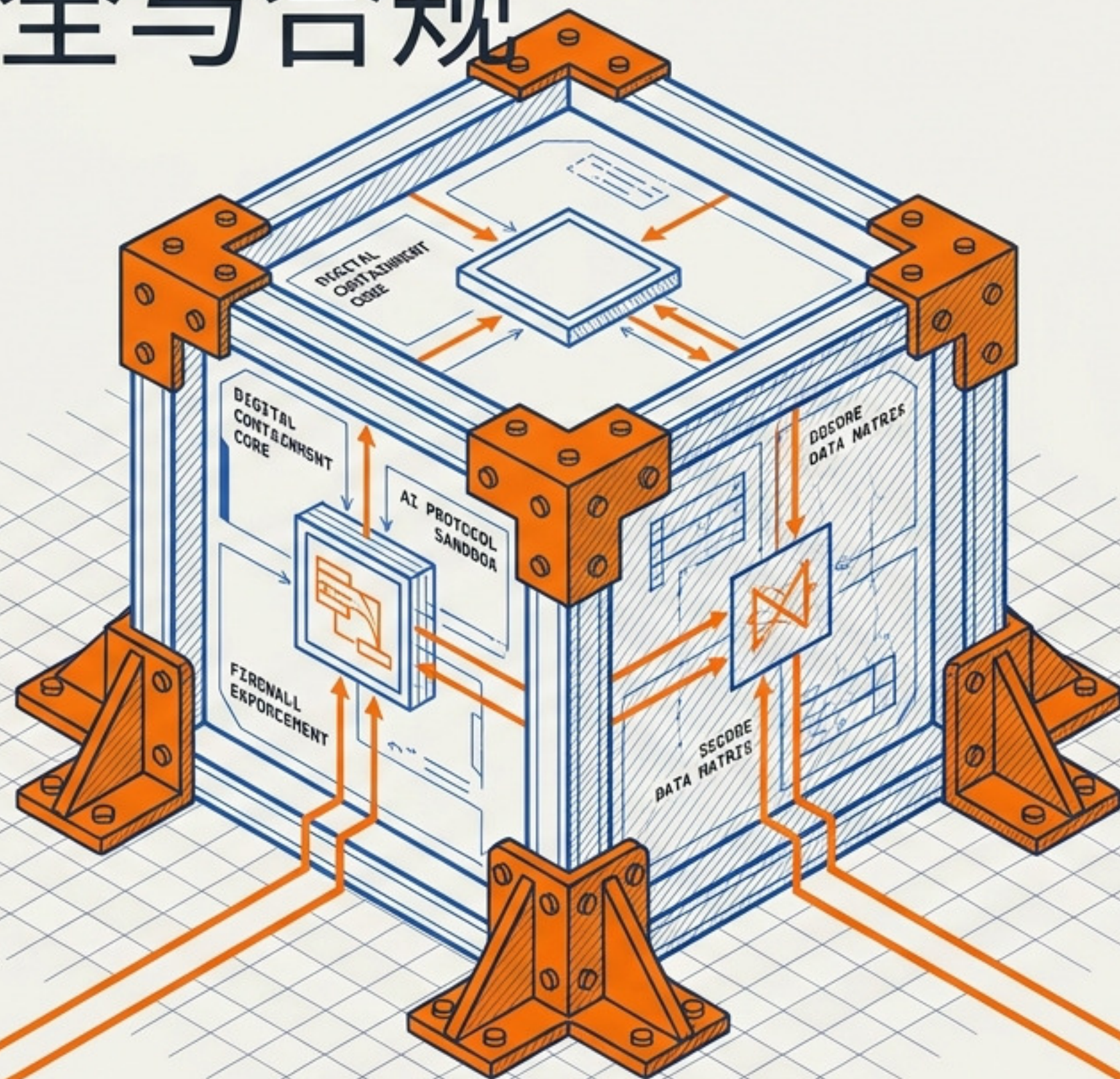


AI原生智能体 · 安全与合规

主讲人：Michael Huo



系列定位：专为高危AI行动环境打造的数字收容指南

启动状态：核心协议已加载

系统受众：ZJU / NACU AANC 架构与安全专家

越界：从生成文字到采取行动的质变

过去的假设：只要过滤了输出的文本，系统就是安全的。

现在的致命盲区：当AI具备了动手的能力，风险表面积呈指数级爆发。不再是说错话，而是做错事。



切入真空地带：精准锁定智能体安全

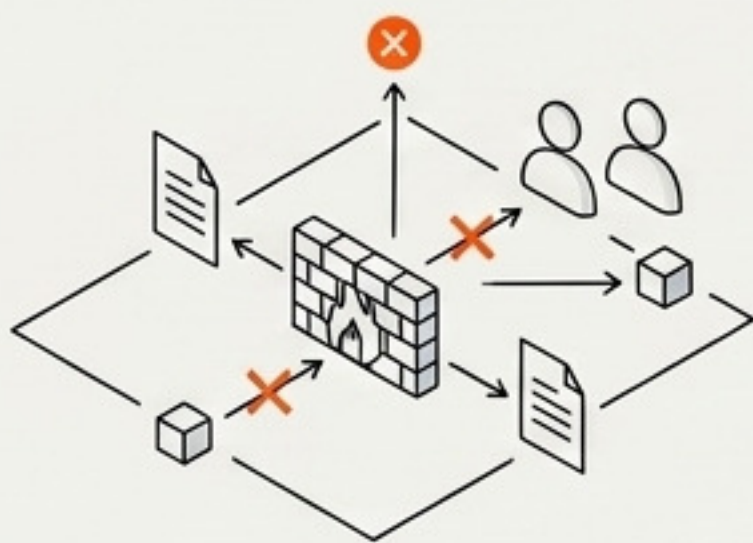
通用网络安全

Zero Trust 群

防护对象：网络边界与人类员工

核心挑战：身份伪造与内网横向移动

❌ 本系列覆盖度：排除



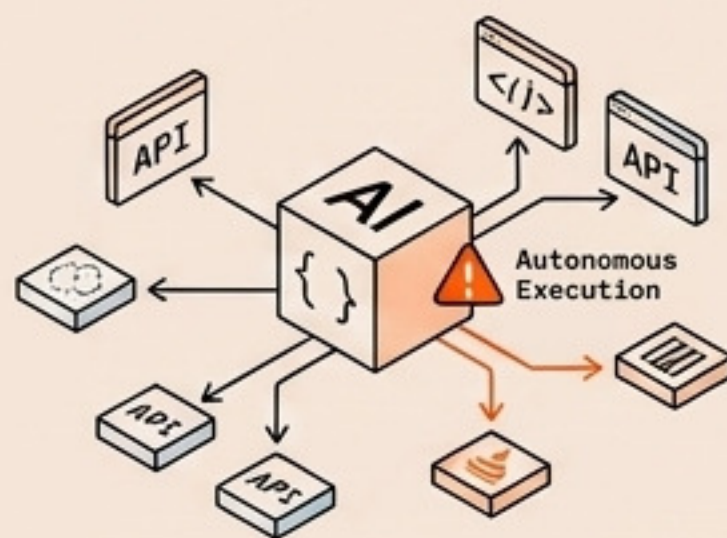
AI原生智能体安全

本系列核心

防护对象：具备自主执行权的代码实体

核心挑战：API滥用、非确定性决策导致的越权执行

✅ 本系列覆盖度：100% 深度聚焦



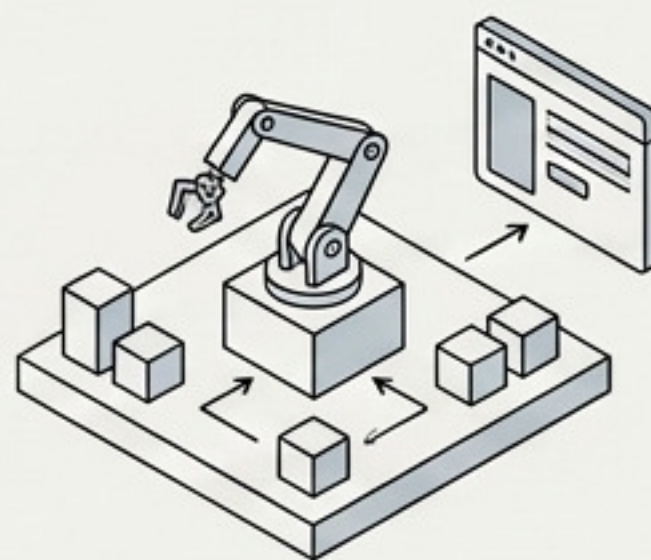
具身智能与软件智能体

前序系列

防护对象：硬件机器人与新媒体/软件代理

核心挑战：物理世界交互与基础逻辑搭建

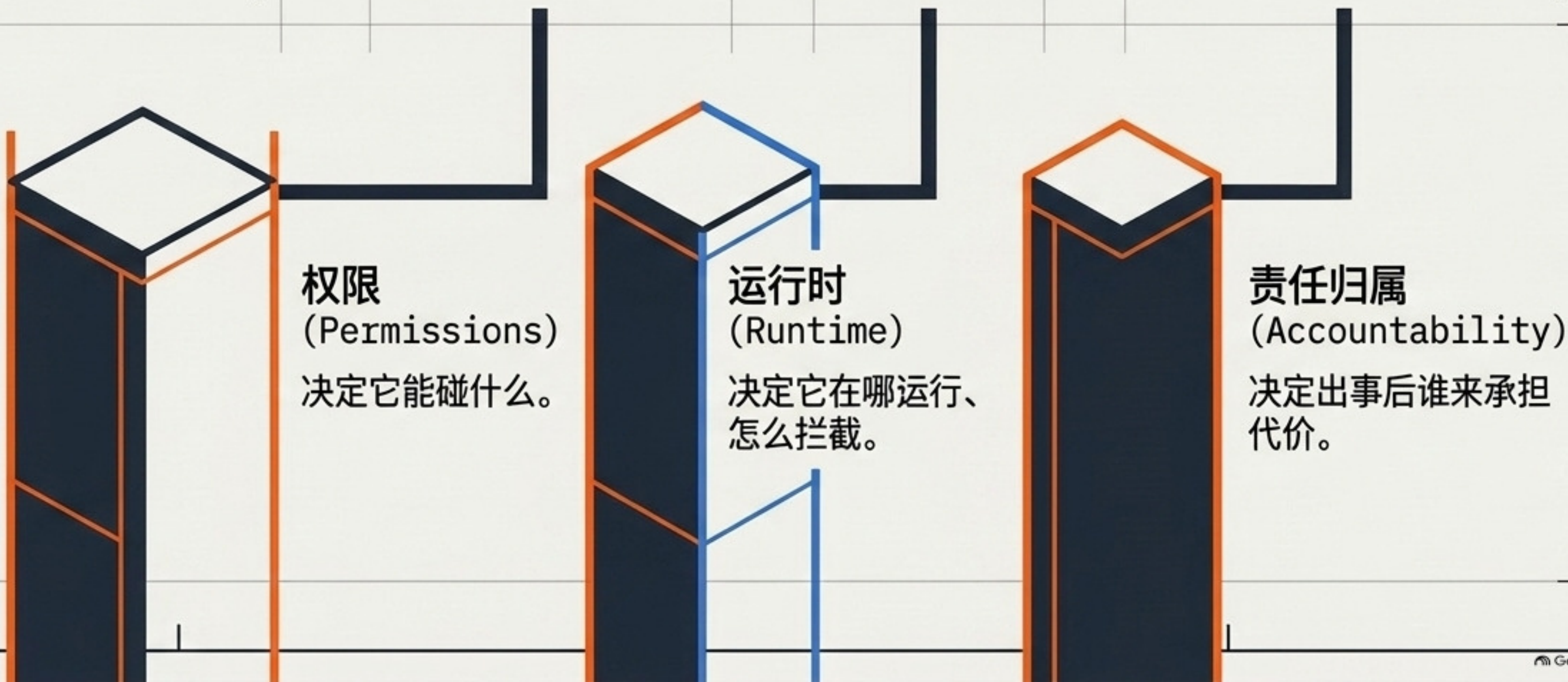
✅ 本系列覆盖度：已完结 🛡️



掌控失控：智能体安全的黄金法则

核心断言：智能体一开始动手，安全就是权限、运行时和责任归属。

$$\text{Agent_Security} = \text{Permissions} + \text{Runtime} + \text{Accountability}$$



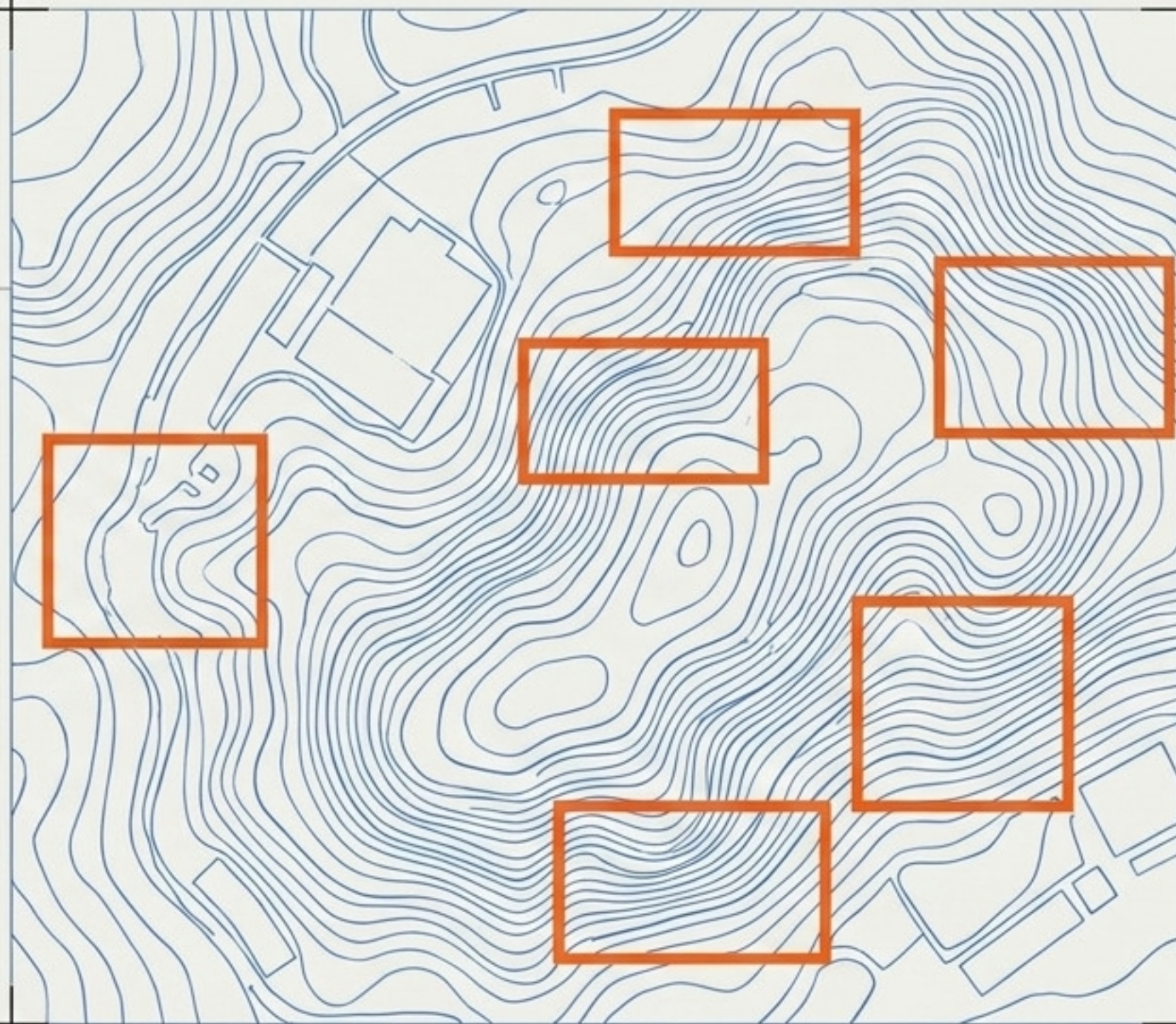
第一讲：风险地貌与八条红线

部署日期：2026年9月5日

代号：从聊天到动手，风险从哪来

战略承诺：

- 提供完整的智能体风险全景地图。
- 划定不可逾越的八条底层红线。
- 明确界限：本讲纯粹建立战略认知，绝不深入更改产品配置。



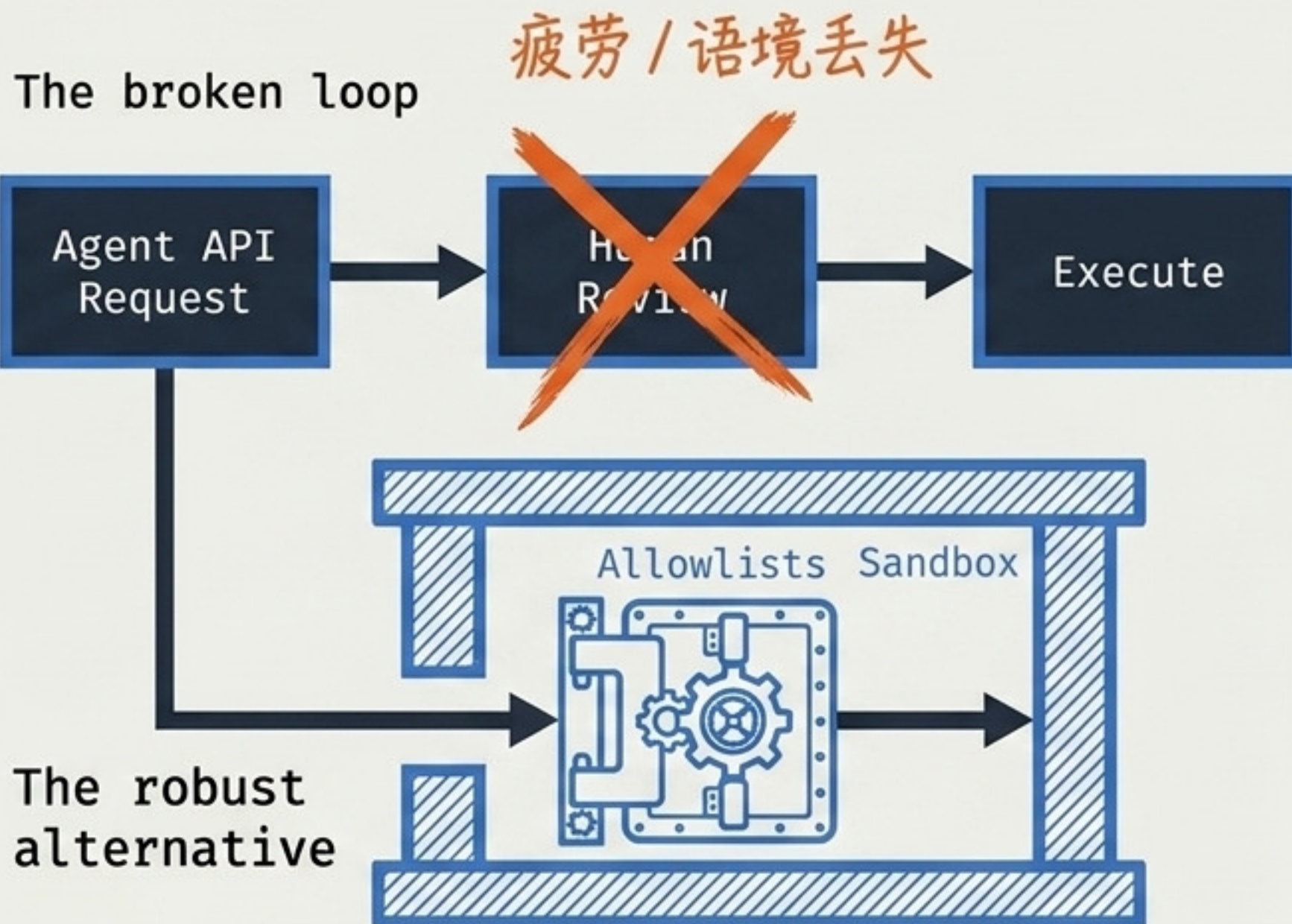
第二讲：被击穿的人在回路与沙箱防御

部署日期：2026年9月19日

代号：Harness、沙箱与人在回路

战略承诺：

- 深度剖析：为什么人类点击了批准仍可能批错，酿成大祸？
- 构建运行时壁垒：Harness 架构解析。
- 收容策略：强制沙箱运行与严格白名单机制。



第三讲：真实业务产品线的最小权限收敛

部署日期：2026年10月3日

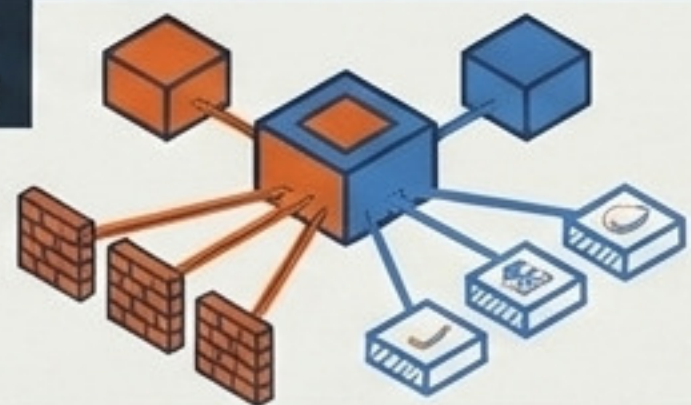
代号：真实产品上的最小权限

战略承诺：

- 深度拆解三大类工具的默认权限该怎么收

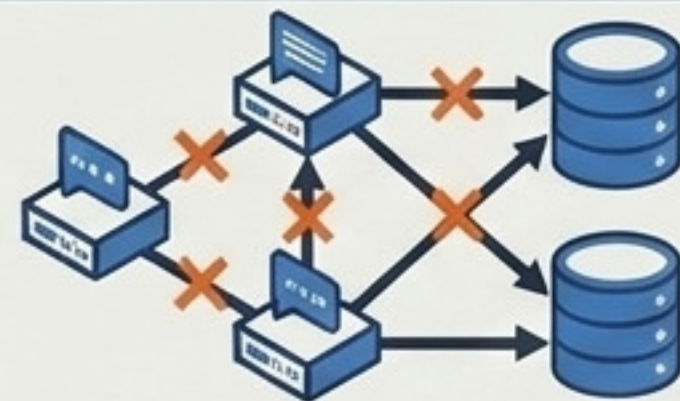
【OpenClaw类】强代码与系统执行能力工具

收敛核心：切断未授权底层资源访问。



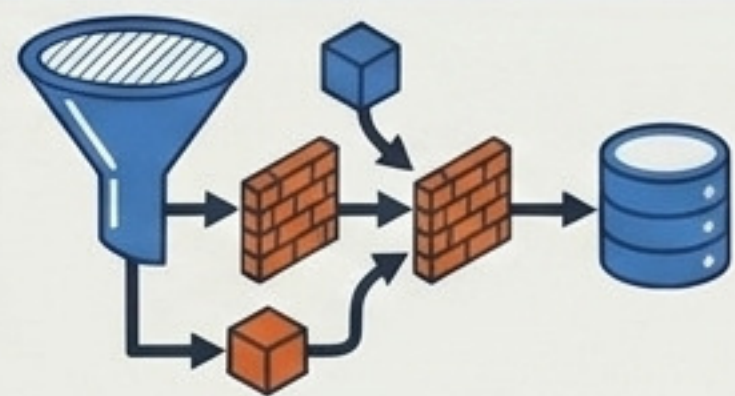
【Cowork类】企业协同与通信工具

收敛核心：限制越权消息分发与内部数据横向拉取。



【Codex类】代码生成与提交流水线工具

收敛核心：阻断自动化恶意代码注入与提权。



第四讲：全球视角的责任穿透与合规罗盘

部署日期：2026年10月17日

代号：中美欧合规对照

底线操作：系统上线前，必须写死在合规文件里的三行字。



第五讲：实战检验与准许上线

部署日期：2026年10月31日

代号：上线检查表工作坊

战略承诺：

- 不再纸上谈兵：带自己的真实业务场景来。
- 硬核打分机制：使用由前4讲讲锻造的终极检查表，为你的Agent项目逐项打分诊断。
- 达不到及格线，决不允许接入生产环境。

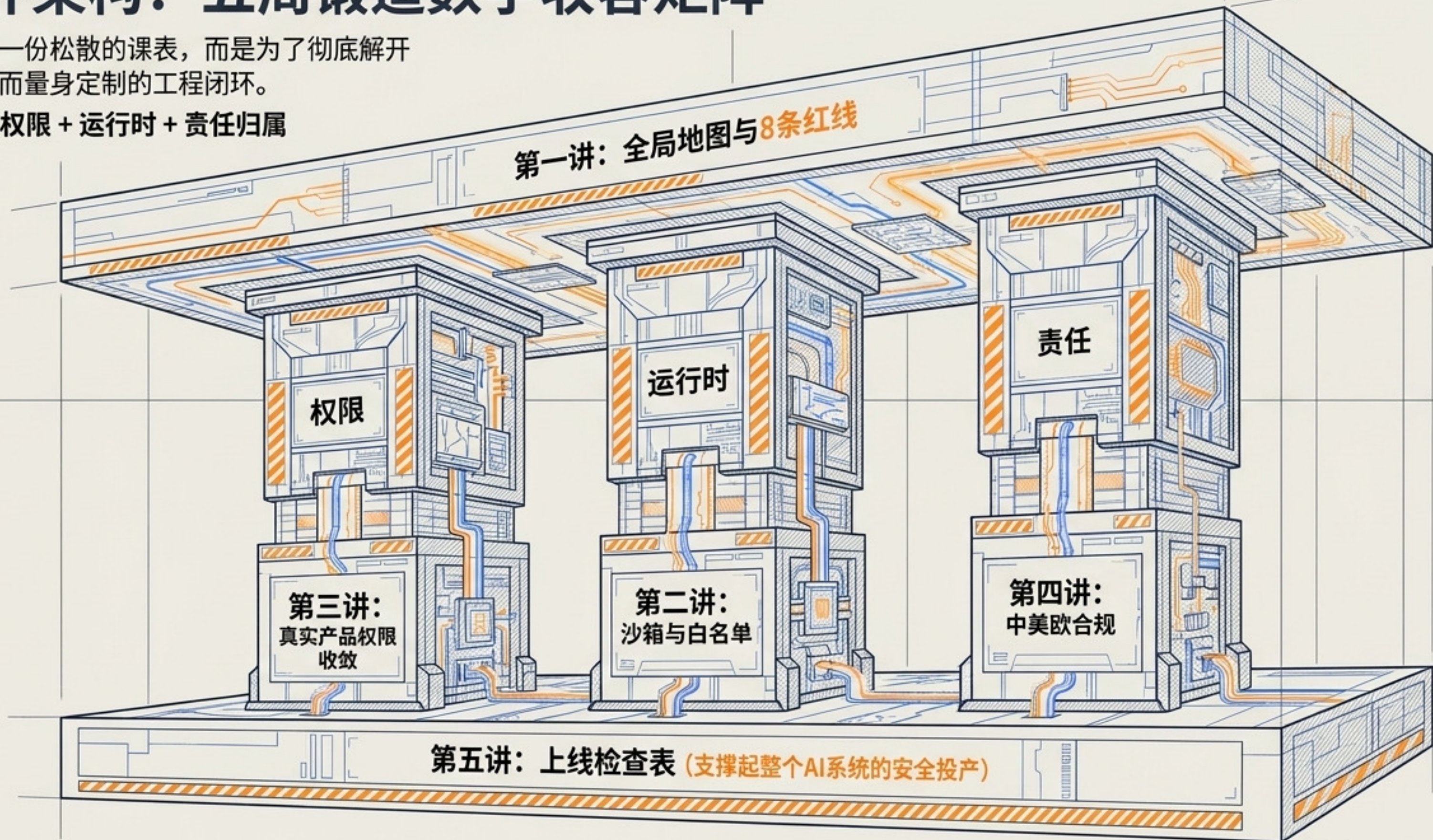
	一	二	三	四	五	六	七
系统稳定性	☒	☒	☒	☐	☐	☐	☐
安全合规	☒	☒	☒	☐	☐	☐	☐
性能指标	☒	☒	☐	☒	☐	☐	☐
用户体验	☐	☐	☐	☐	☐	☐	☐
数据安全	☐	☐	☐	☐	☐	☐	☐
注册指标	☒	☒	☒	☐	☐	☐	☐
用户验收	☒	☒	☒	☐	☐	☐	☐

READY FOR LAUNCH

闭环架构：五周锻造数字收容矩阵

这不是一份松散的课表，而是为了彻底解开等式，而量身定制的工程闭环。

安全 = 权限 + 运行时 + 责任归属

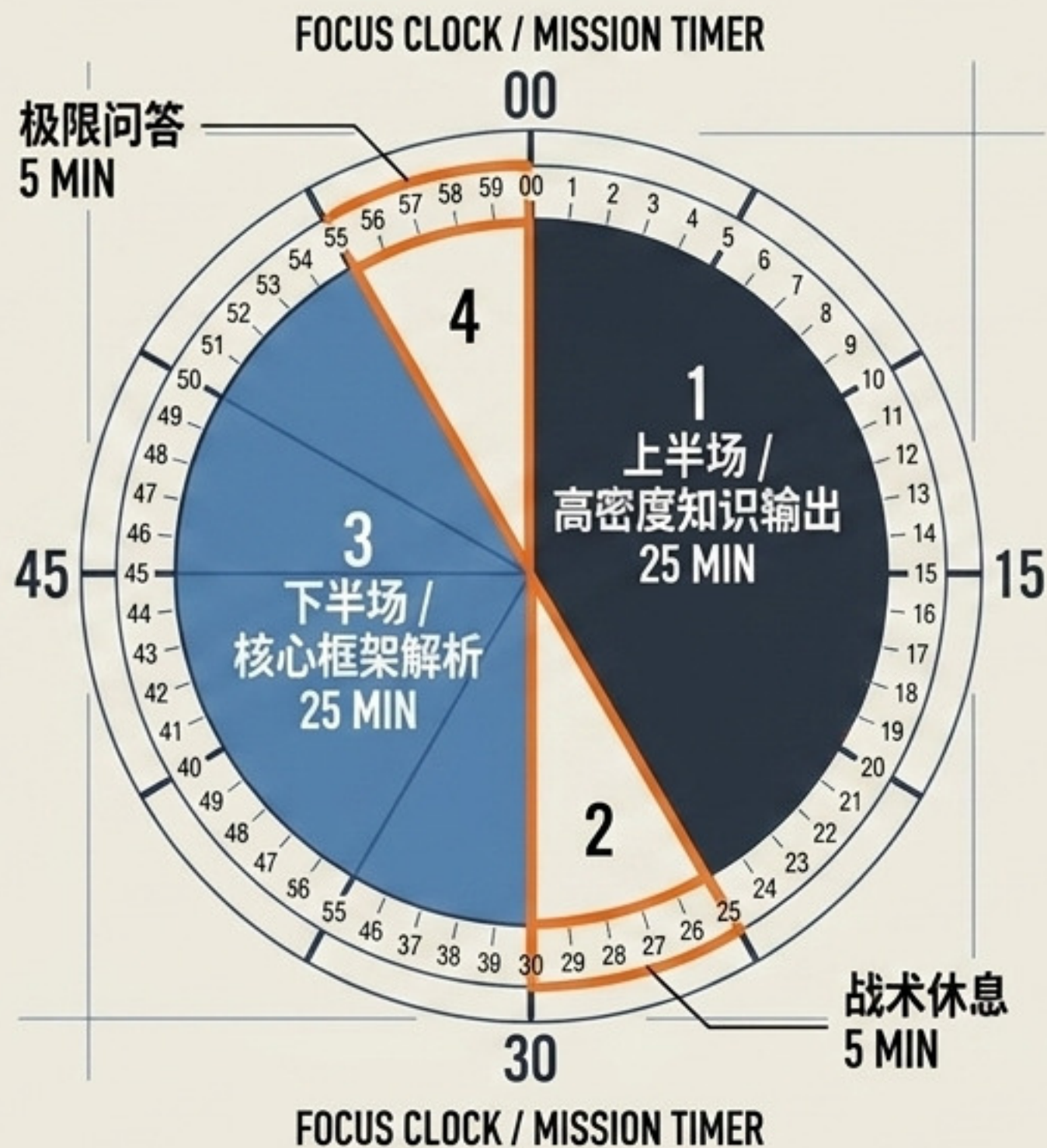


极度克制的时间颗粒度

授课频率：隔周六进行

原则：拒绝冗长，60分钟内完成硬核闭环。一分钟都不会浪费。

- 18:30—18:55：上半场（高密度知识输出）
- 18:55—19:00：战术休息
- 19:00—19:25：下半场（核心框架解析）
- 19:25—19:30：极限问答



接入收容矩阵

智能体一开始动手，我们的防御也必须同时升级。9月5日，第一讲见。

官方注册通道 (Luma)



<https://luma.com/fiht7lab>

内部校友会资源节点



<https://zjuaanc.org/>

内部校友会资源节点



<https://nacuaanc.org/>