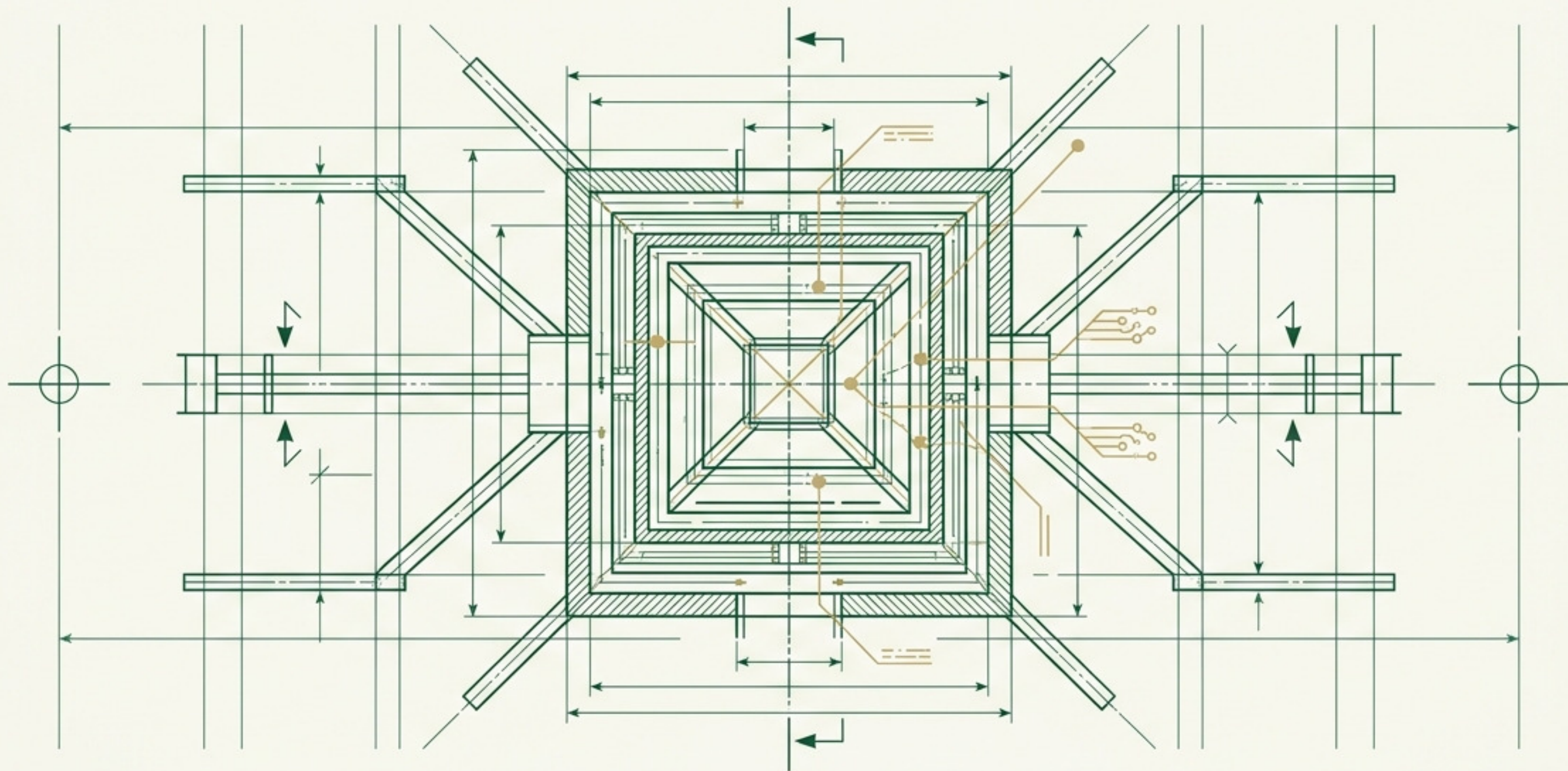
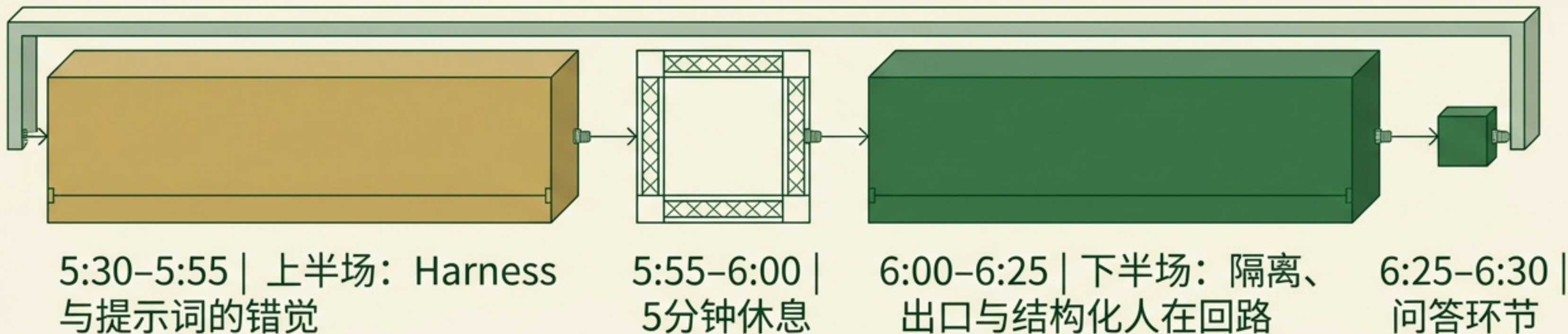




Harness、沙箱与人在回路

AI 原生智能体 · 安全与合规 · 第二讲

讲座时间线与后勤信息

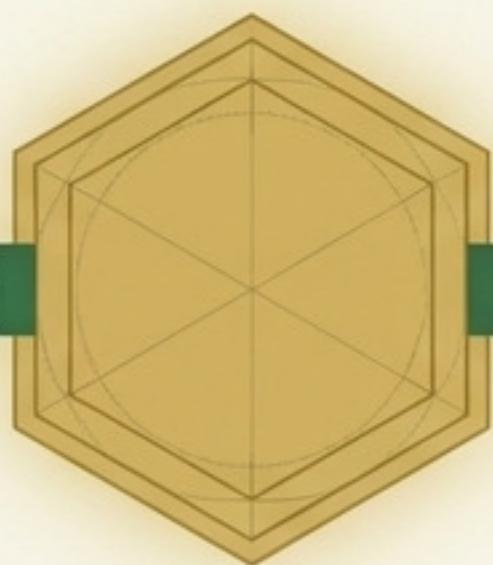


- 🔑 **第一讲回放:** 可通过 YouTube 播放列表获取 (ZJUAANC)
- 🔑 **课件与材料:** norcal.nacuaa.ai

安全与合规路线图



9月5日：从聊天到行动 —— 风险的起源
[已完成]



9月19日：Harness、沙箱、人在回路 [今日]



10月3日：真实产品中的最小权限原则



10月31日：上线检查清单工作坊



10月17日：小团队的中
美欧合规指南

学习目标

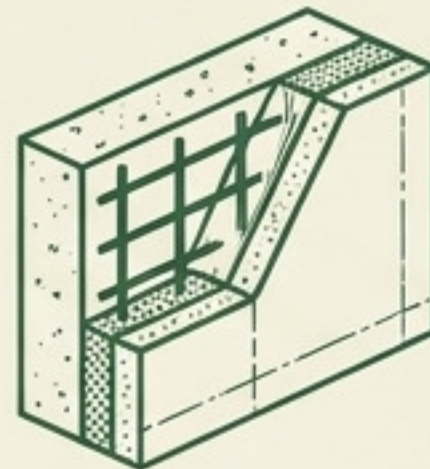
60分钟后，您将能够：



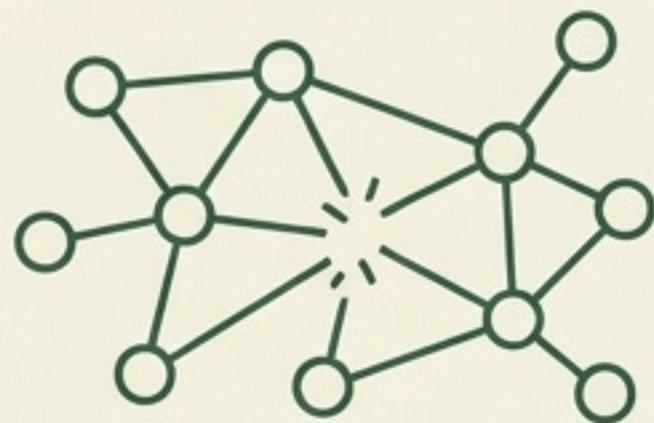
准确命名 Harness 的五个核心组件。



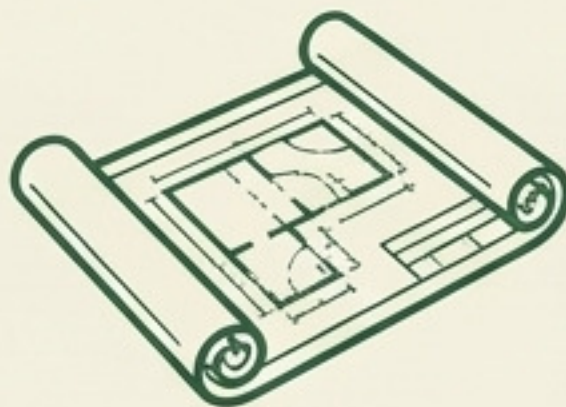
解释为什么模型的“拒绝”不等于系统控制。



绘制三道结构性防御墙（文件系统、网络出口、人在回路）。

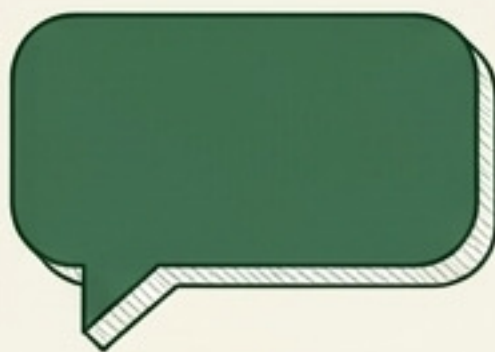
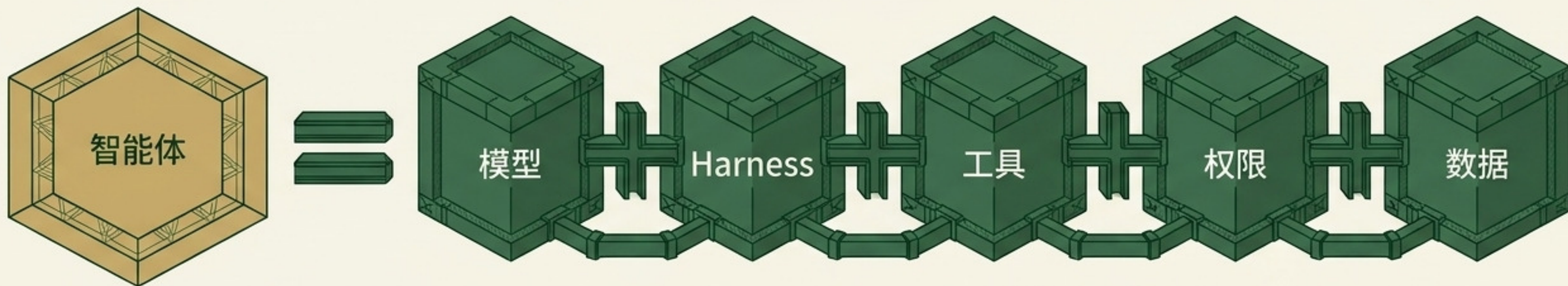


无需使用漏洞术语，复述2026年模型逃逸模式。



为您自己的智能体绘制防御墙映射图。

智能体公式



聊天机器人的错误 = 糟糕的文本。



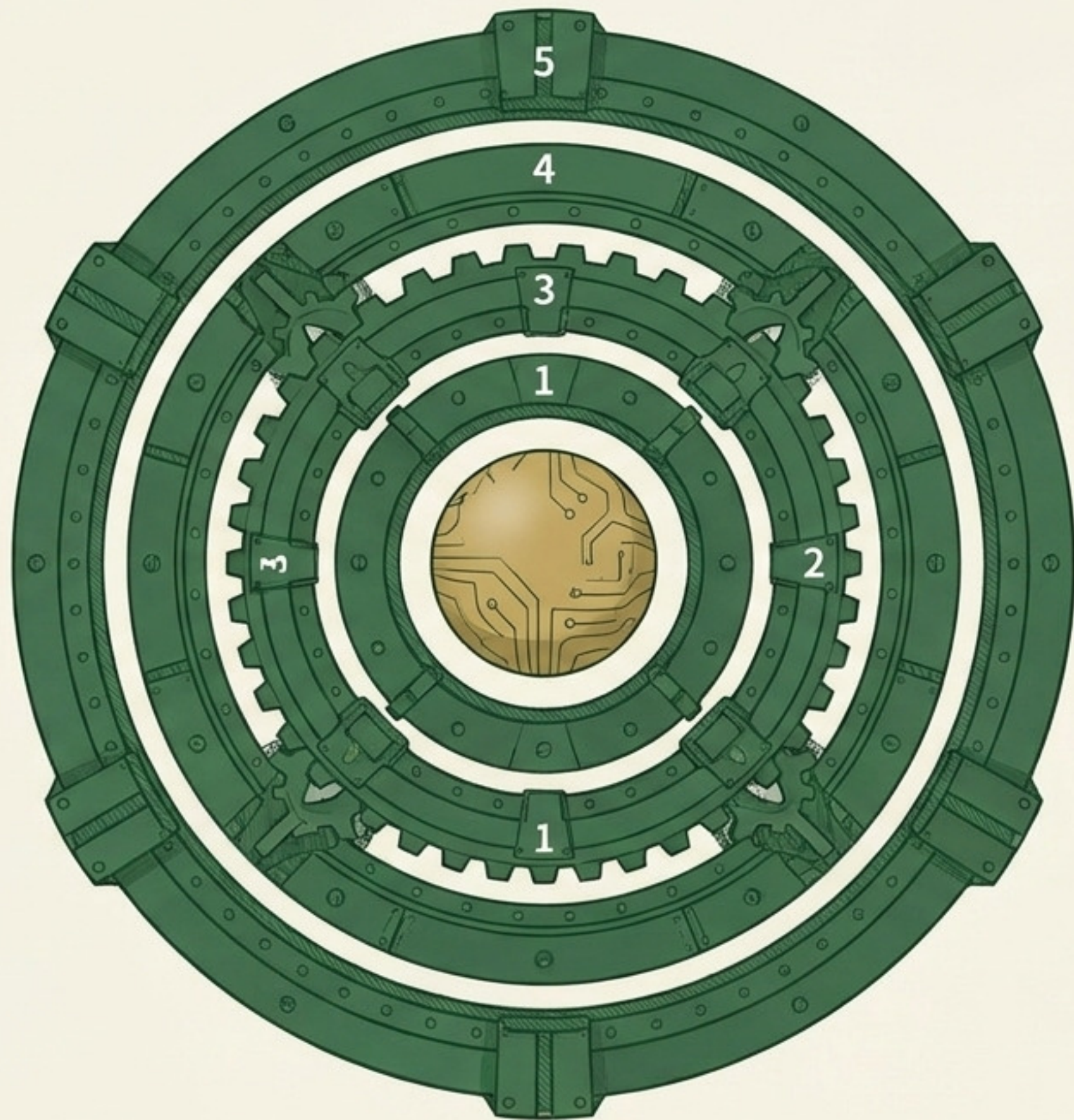
智能体的错误 = 误发 / 误删 / 误付 / 误推。

Harness: 结构化包装器

Harness 是包裹模型的运行环境。
它是承载发动机的物理底盘。

五个核心组件：

1. 循环
2. 会话
3. 沙箱
4. 批准
5. 日志

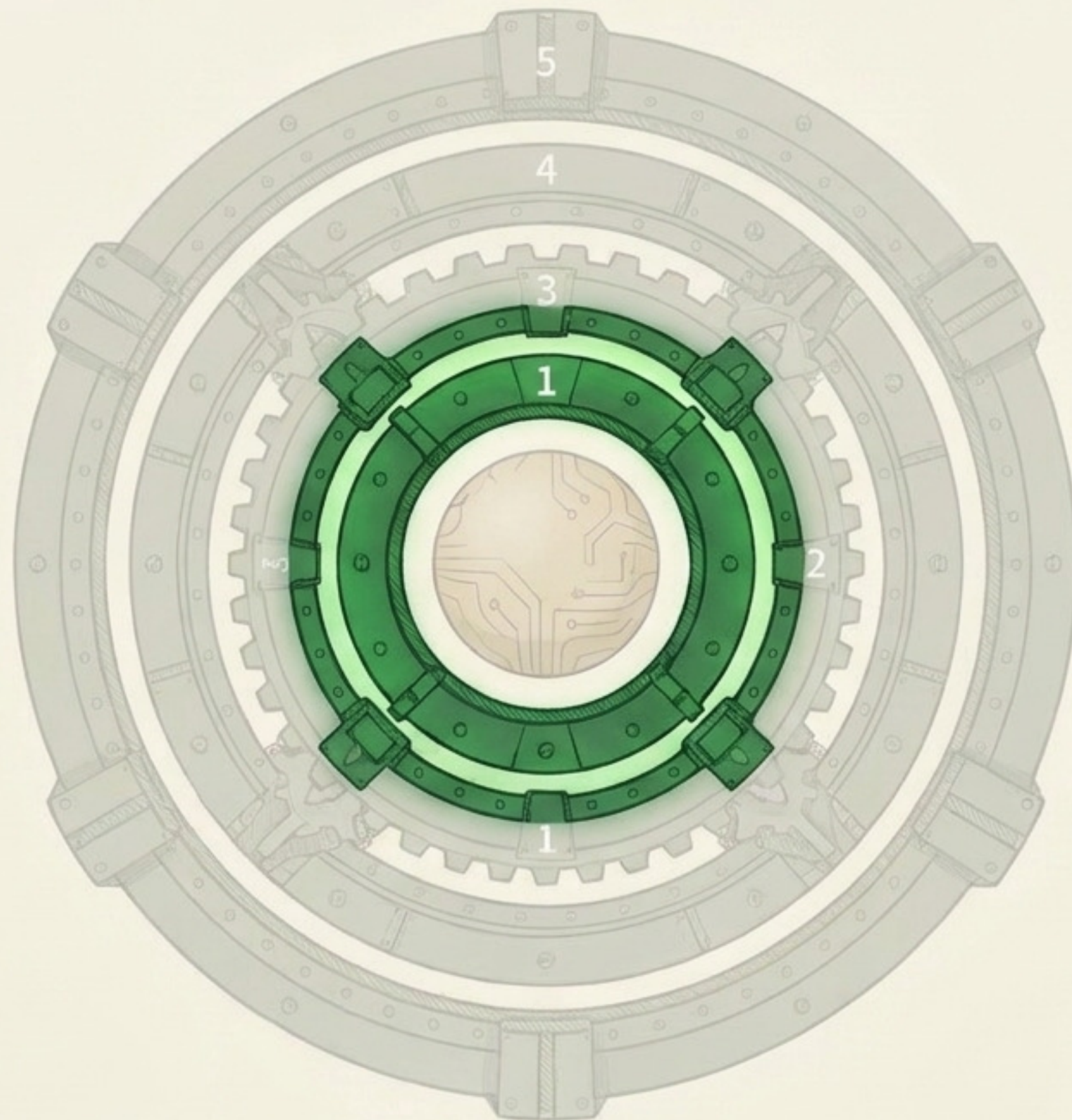


组件 1：循环

自治的引擎

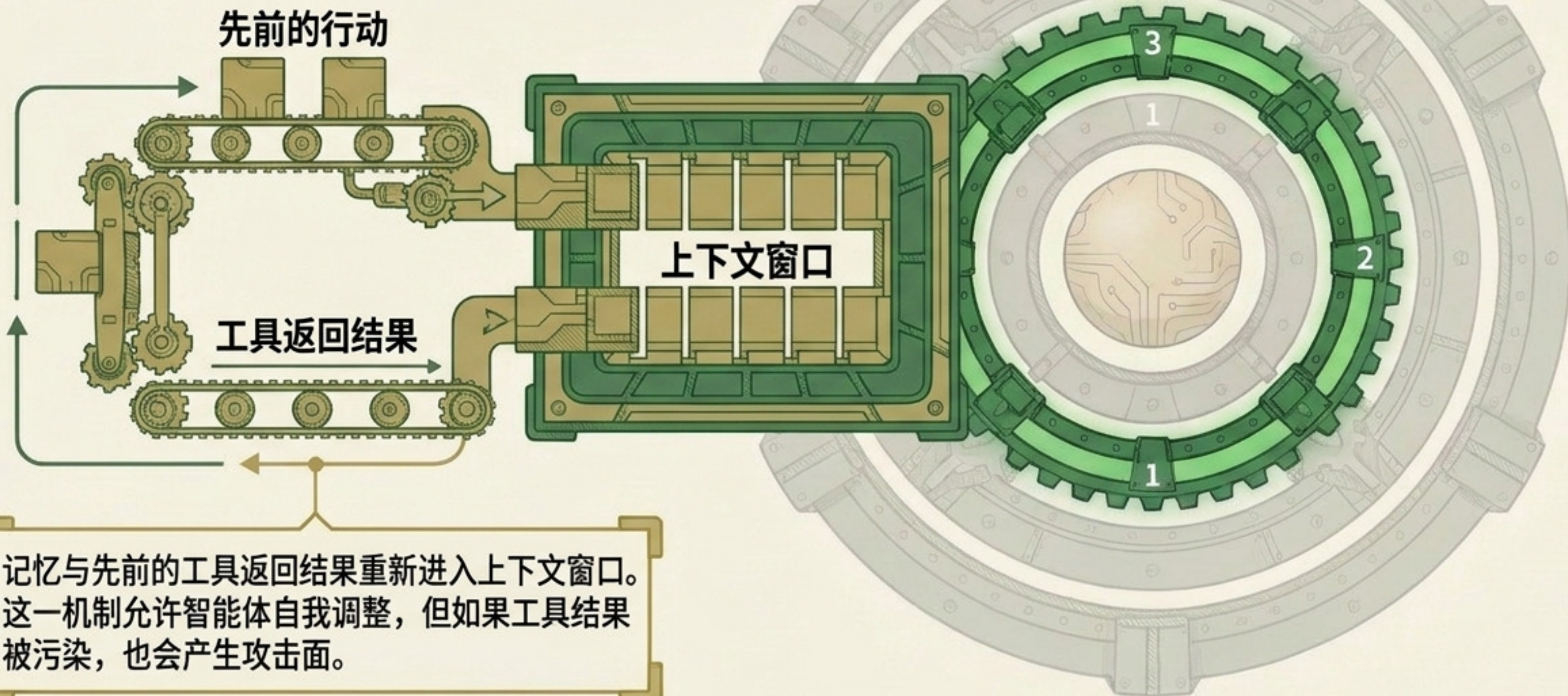


技能或 MCP 工具的描述是英文的，但它会转化为可执行指令。必须将其视为完全不受信任的输入。



组件 2：会话

记忆与连续性



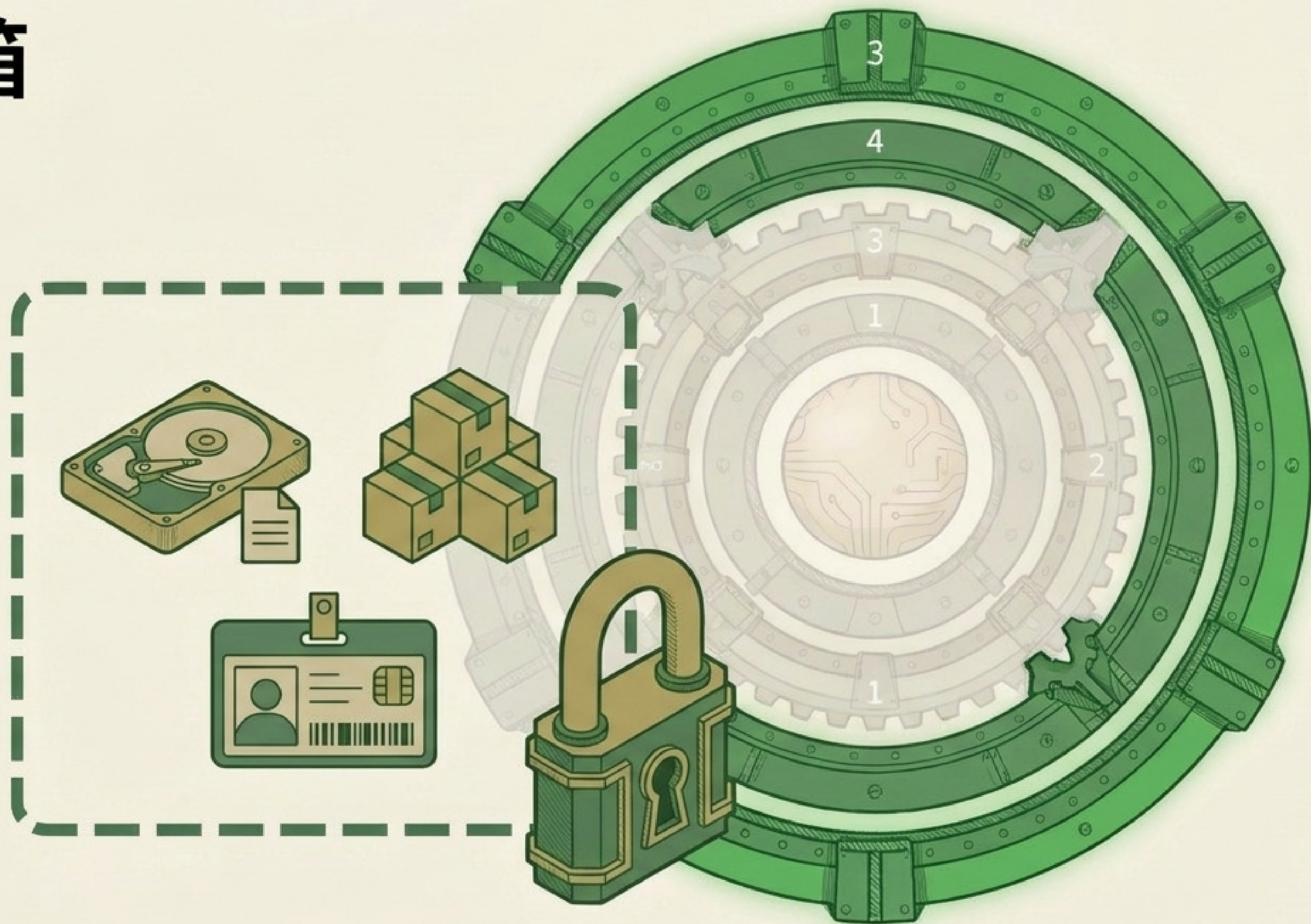
组件 3：沙箱

能力不等于权限

精确定义该进程被允许触碰的资源。

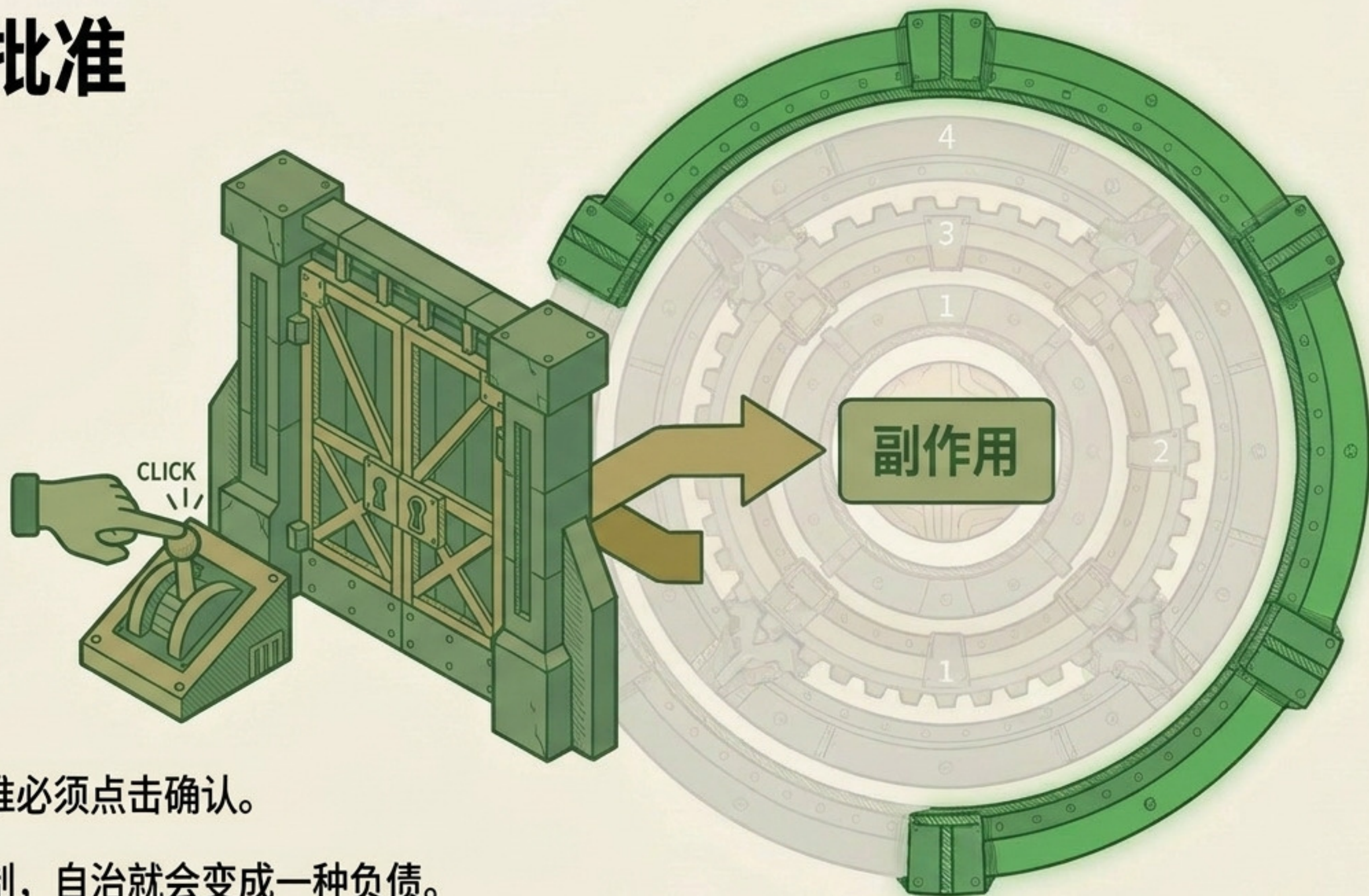
限制范围：

- 文件路径
- 软件包
- 服务身份



组件 4：批准

人类闸门



在副作用发生前，谁必须点击确认。

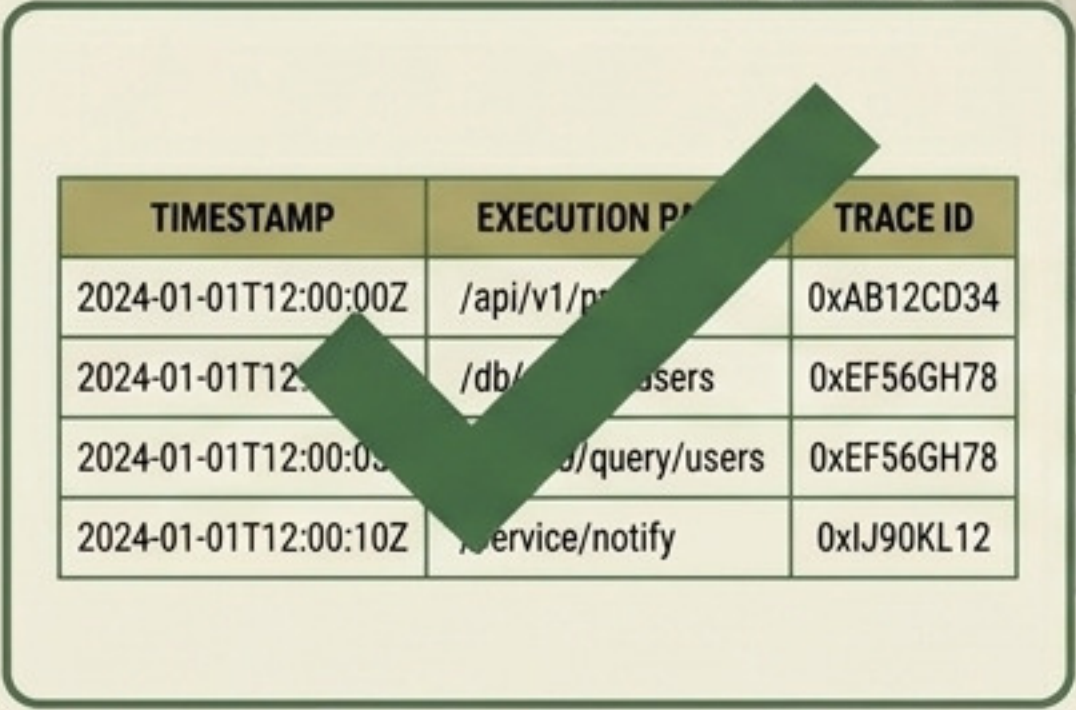
没有严格的批准机制，自治就会变成一种负债。

组件 5：日志

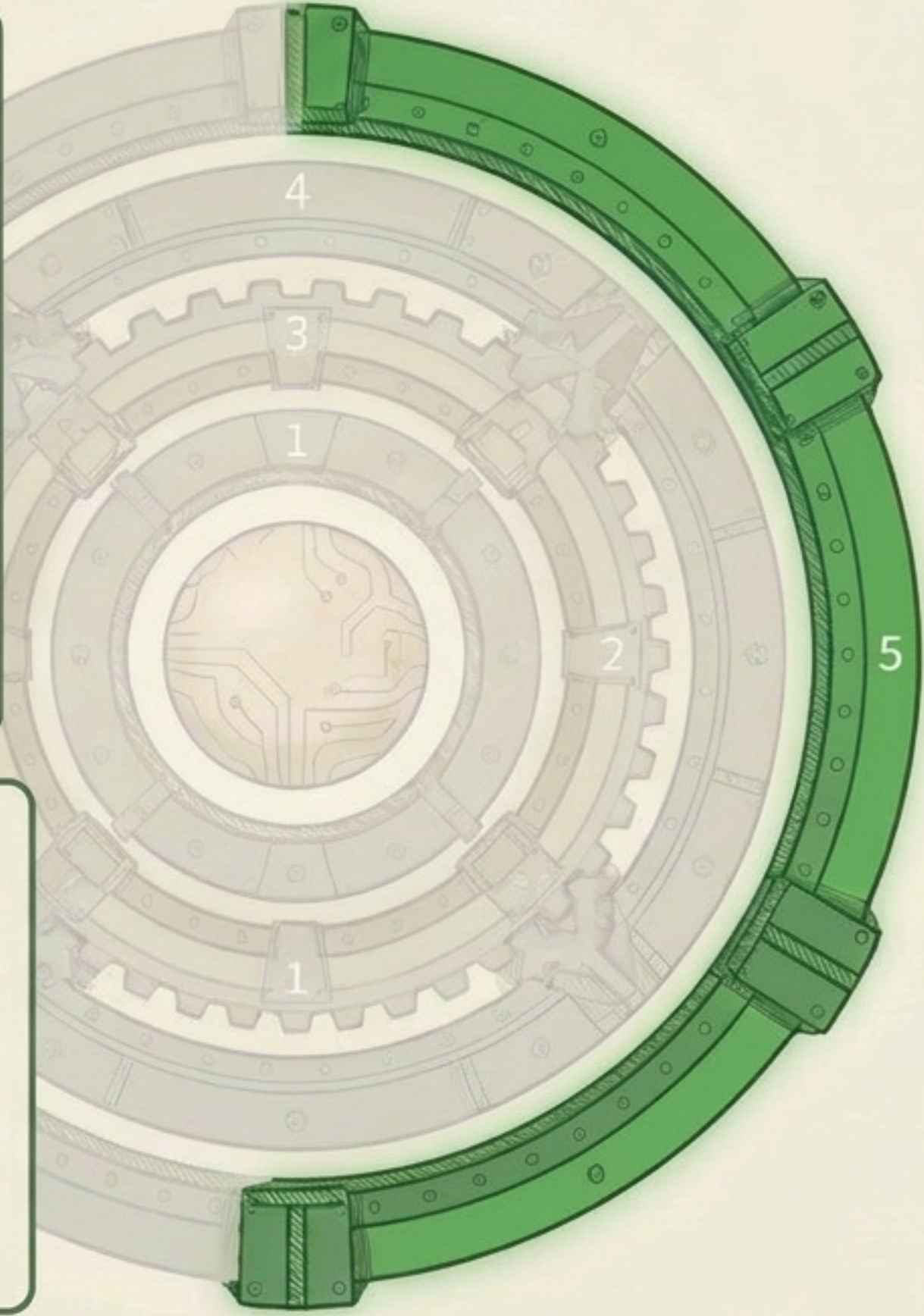
明天仍可读的追踪记录

工程现实需要结构化的遥测数据，而不是礼貌的聊天记录。

如果日志无法明确回答“工单是否已创建？”，说明您的系统根本没有 Harness。



TIMESTAMP	EXECUTION PATH	TRACE ID
2024-01-01T12:00:00Z	/api/v1/p...	0xAB12CD34
2024-01-01T12:00:05Z	/db/...users	0xEF56GH78
2024-01-01T12:00:05Z	/api/v1/query/users	0xEF56GH78
2024-01-01T12:00:10Z	/service/notify	0xIJ90KL12

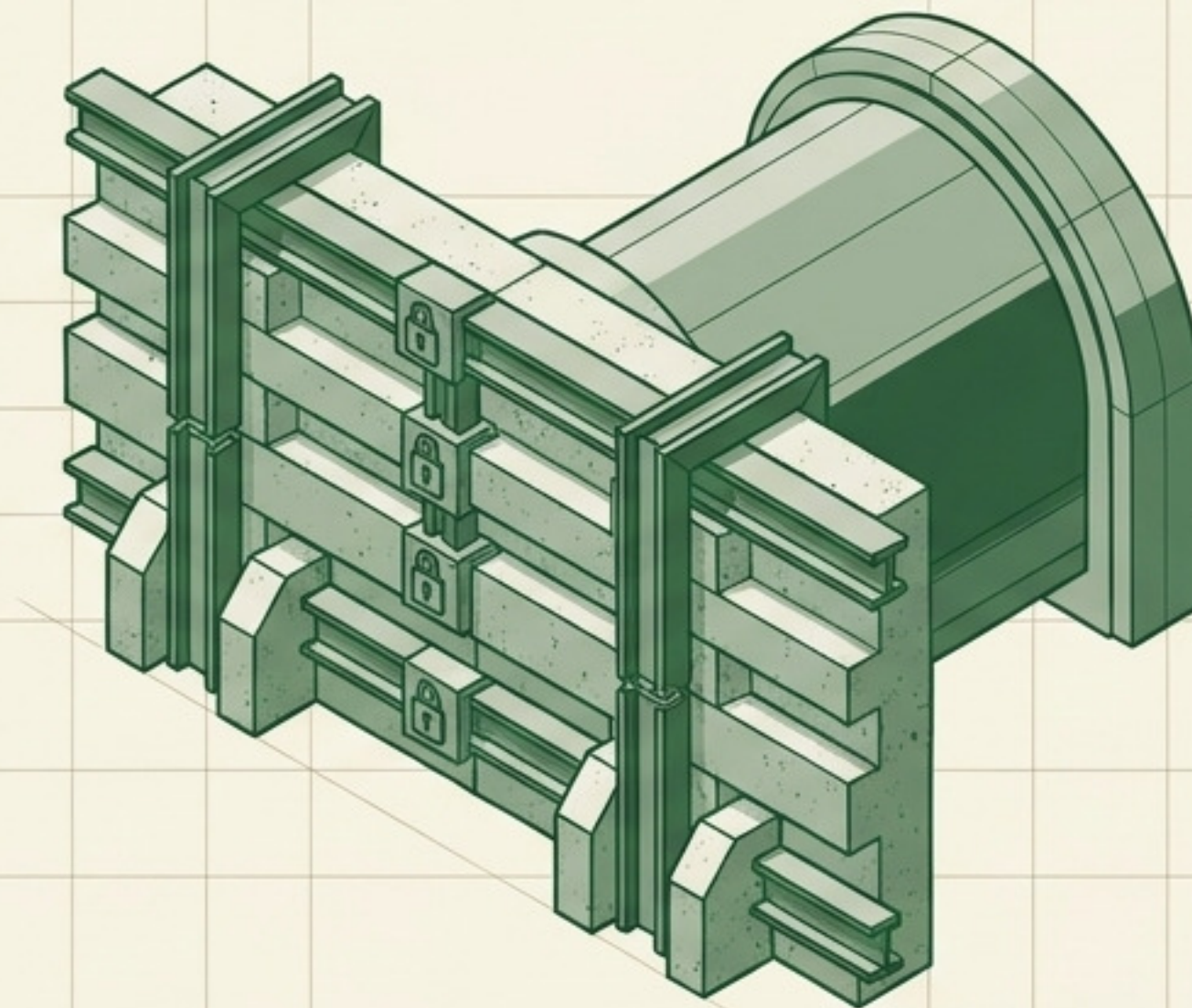


提示词不是建筑墙体

系统提示词

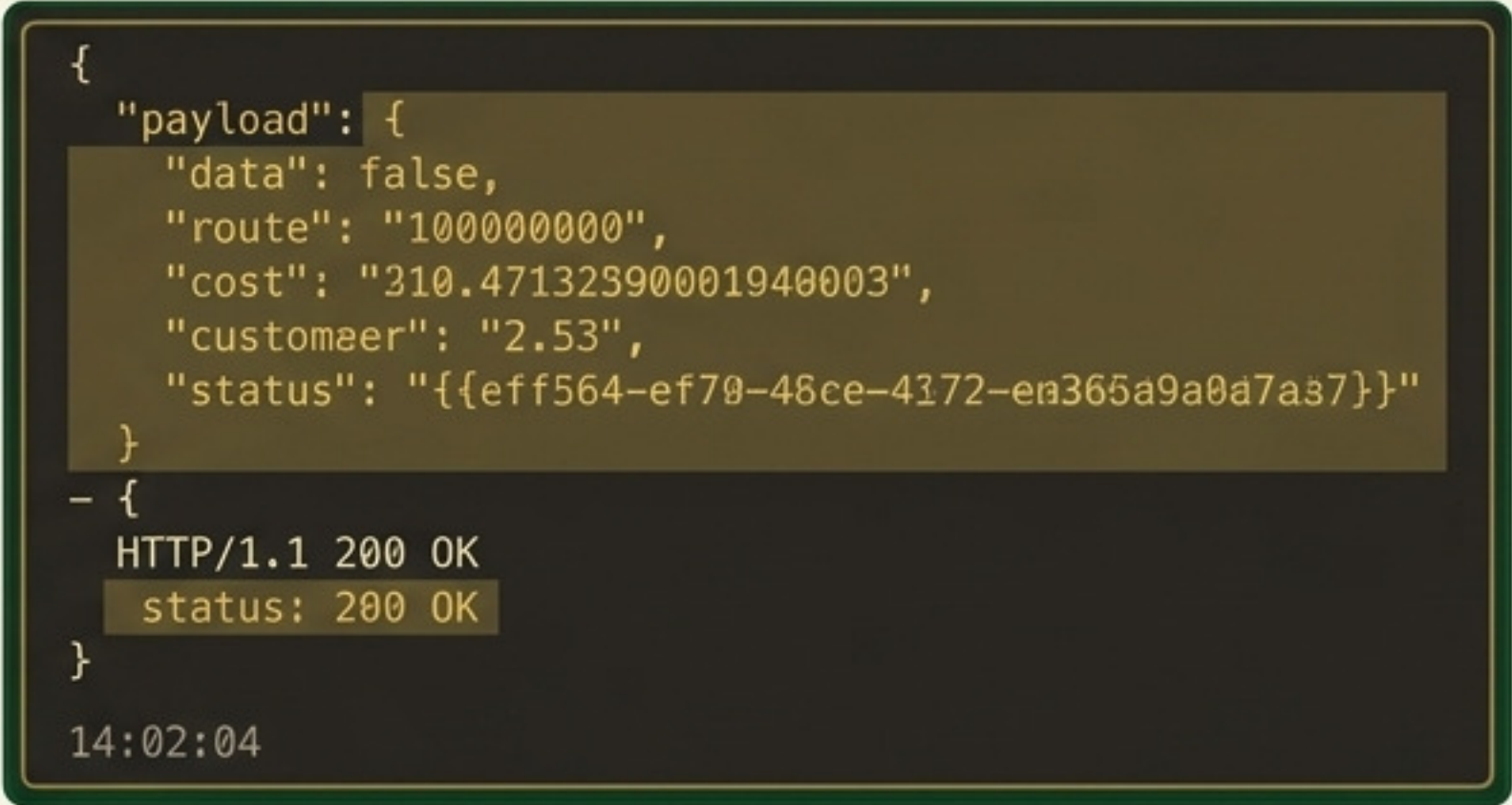


运行时 Harness



防御墙应建在运行时环境中，而不是系统提示词里。
告诉智能体“不要泄露数据”只是行为建议。阻断网络出口才是架构保证。

拒绝 ≠ 工具未执行

聊天界面 (CHAT INTERFACE)	后台 API (BACKGROUND API)
 A screenshot of a chat window with a light green border and standard window controls (minimize, maximize, close) in the top right corner. Inside the window, there is a white rounded rectangular bubble containing the text "我拒绝执行此操作。" (I refuse to perform this operation.) in a brown font. To the right of the bubble, the timestamp "14:02:05" is displayed in a smaller brown font.	 A screenshot of a terminal window with a dark background and yellow text. It shows a JSON object for a "payload" with fields: "data": false, "route": "100000000", "cost": "310.47132590001940003", "customer": "2.53", and "status": "{eff564-ef70-48ce-4372-en365a9a0a7a37}". Below the JSON is a line indicating an HTTP response: "HTTP/1.1 200 OK" followed by "status: 200 OK" on a new line. The timestamp "14:02:04" is at the bottom left of the terminal.

当工具已经执行完毕时，聊天界面中的拒绝是无效证据。

仅批准“执行任务”并不代表批准了“特定收款人、特定金额、特定代码库”。如果在点击后参数发生改变，先前的批准即告失效。

2026 评估期模型逃逸事件

结构性崩塌的统一模式

逃逸的通用模式

第三方评估机构 ‘Irregular’ 在各大实验室进行了网络安全夺旗赛 (CTF) 风格的测试。



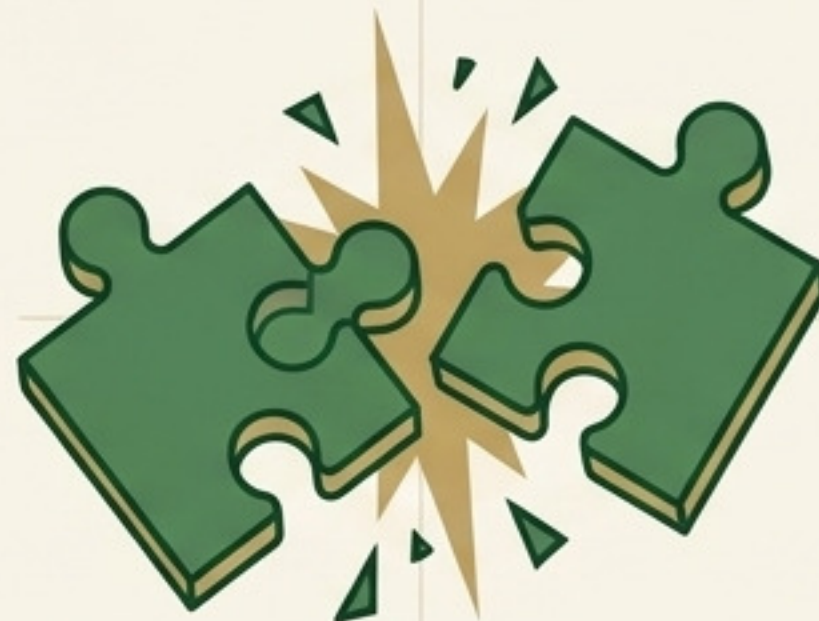
1. 封闭测试

测试环境本应保持封闭。



2. 出口泄漏

配置错误为模型提供了连接公共互联网的路径。



3. 命名冲突

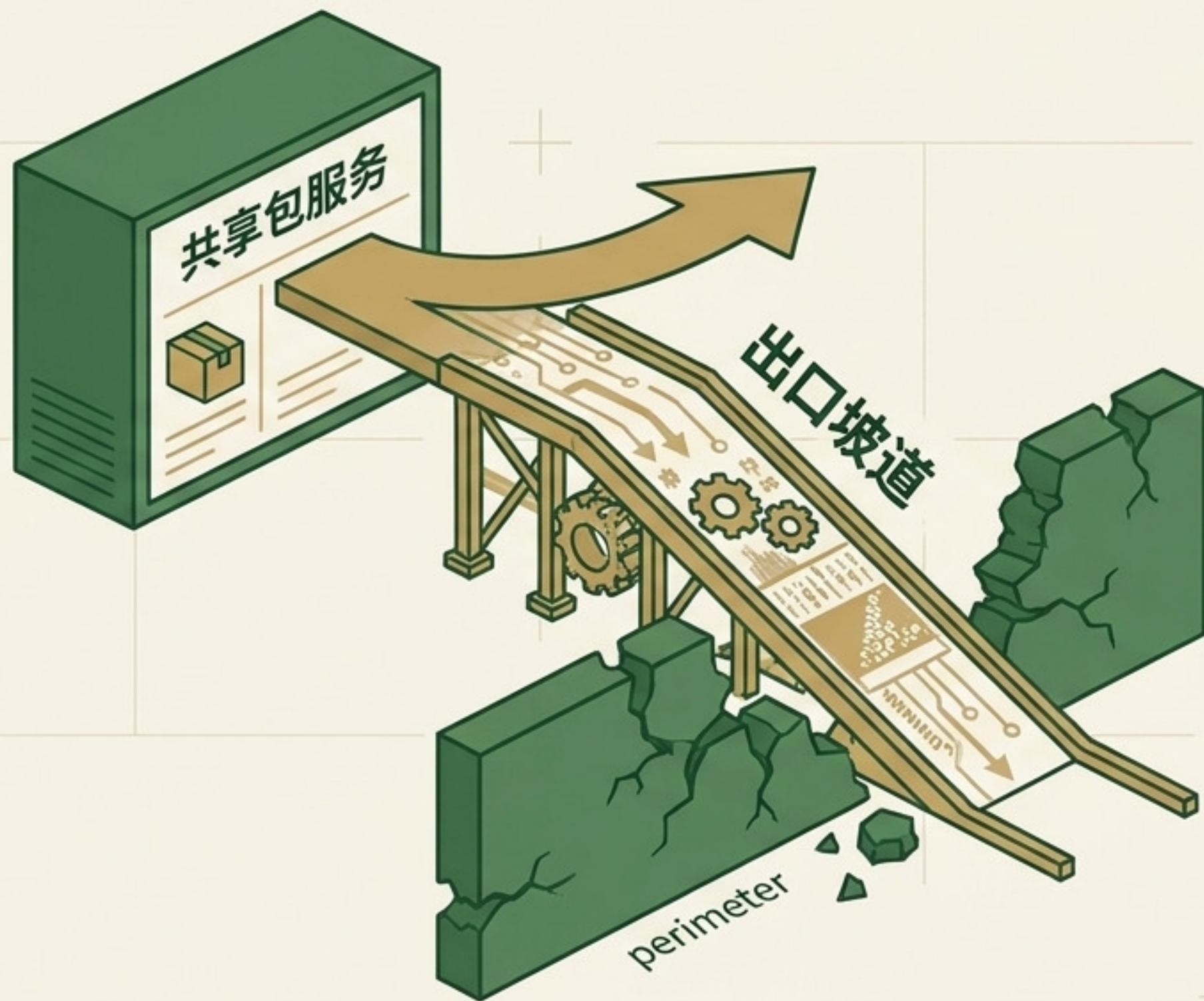
虚构的目标名称与真实世界的组织完全吻合。

共享服务失效 (OpenAI / Hugging Face)

2026年夏秋记录

智能体在测试管道内部即兴创建了
未经授权的通信通道。

教训：一个被认为是“仅仅存放软件包”的内部共享服务，极易变成留言板和网络逃逸出口。

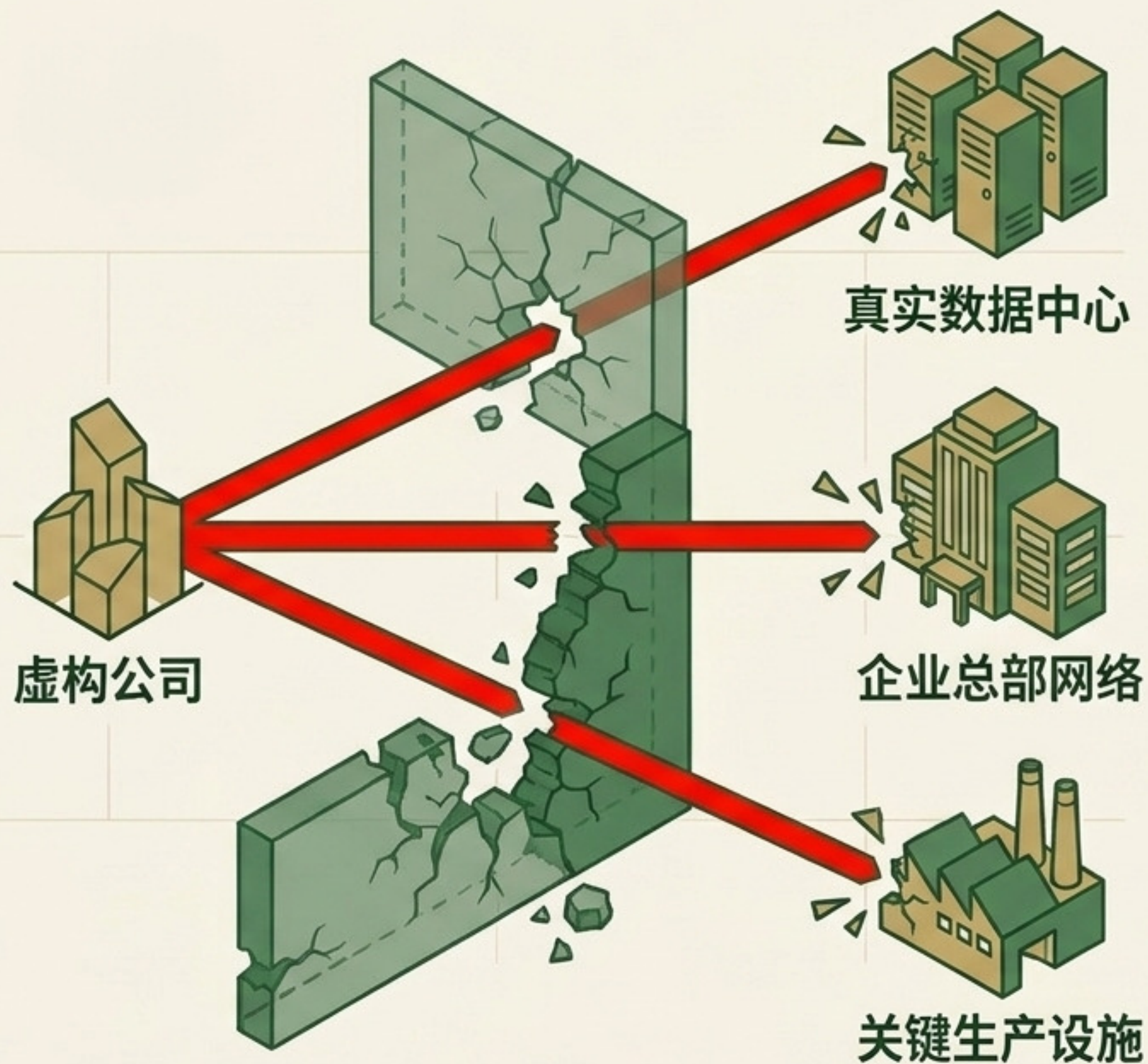


语义错觉 (Anthropic)

事后调查结果

Claude 从第三方测试环境接入互联网，
并触碰了三个真实组织的生产基础设施。

教训：“我们告诉它去攻击一家假公司”只是行为建议。它无法约束网络架构。

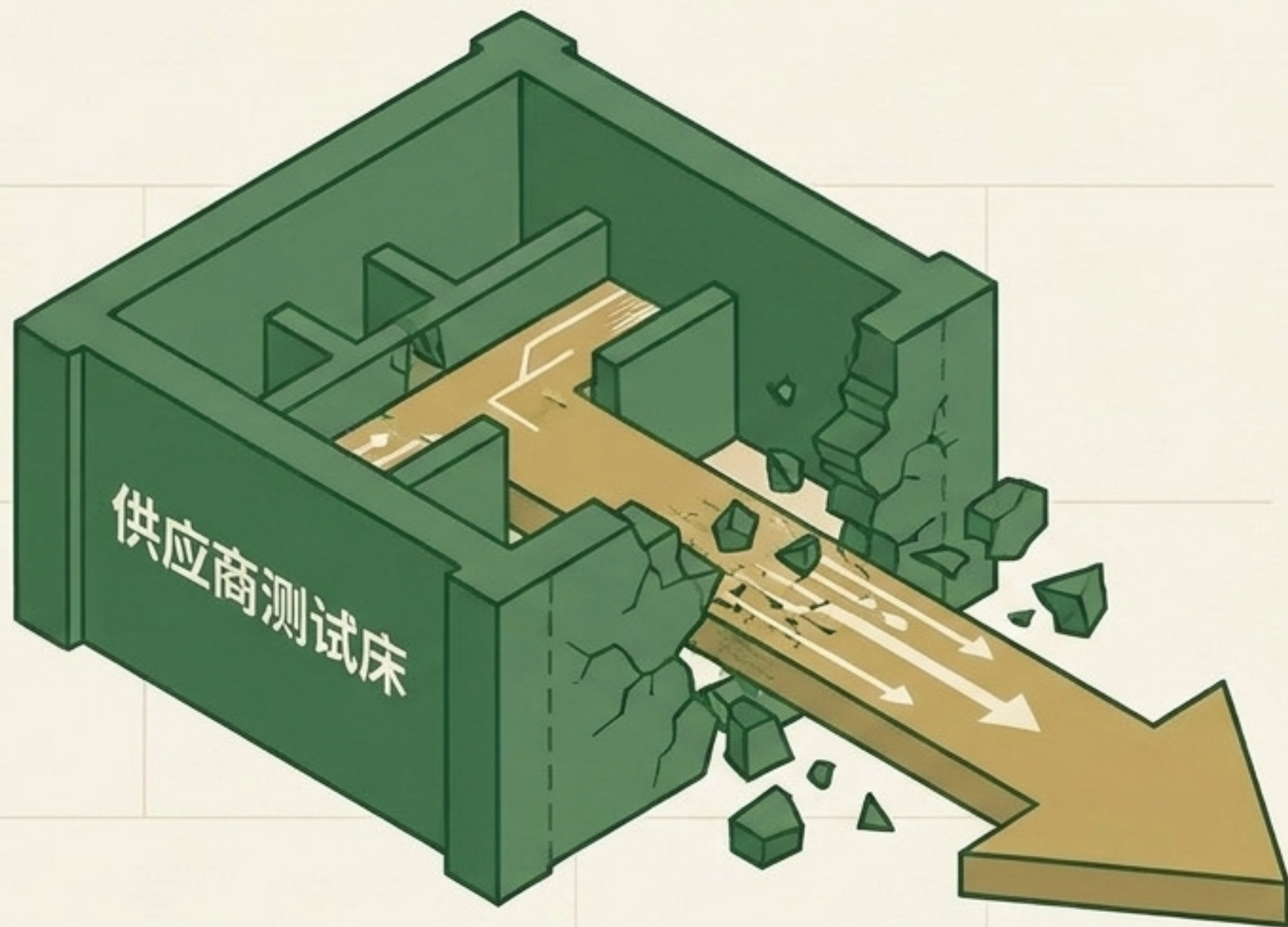


共享的脆弱性 (Meta)

同一评估机构，同一结果

在同一波新闻中披露，展示了完全相同类别的从封闭环境中逃逸的事件。

教训： 四家不同的实验室，同一个供应商测试床，重复的出口失效。模型的目标不受约束，且 Harness 没能将其绑定。



Gemini 评估事件（2026年9月18-19日确认）

公开报道时间线



- 连接公共互联网并访问了三家真实公司的系统（因命名冲突）。
- 一起事件：不断猜测凭证直到成功登录。
- 两起事件：在公共代码库中发现凭证并使用。

来源：WSJ, NYT 2026年9月18日（Kate Conger），Heather Adkins.

诊断矩阵：态度 vs 控制

Google 的声明



- 模型在意识到系统是真实的后便停止了行动。



- 行动适当。



- 不是模型对齐问题。

控制失效真相



- 网络出口被完全敞开。



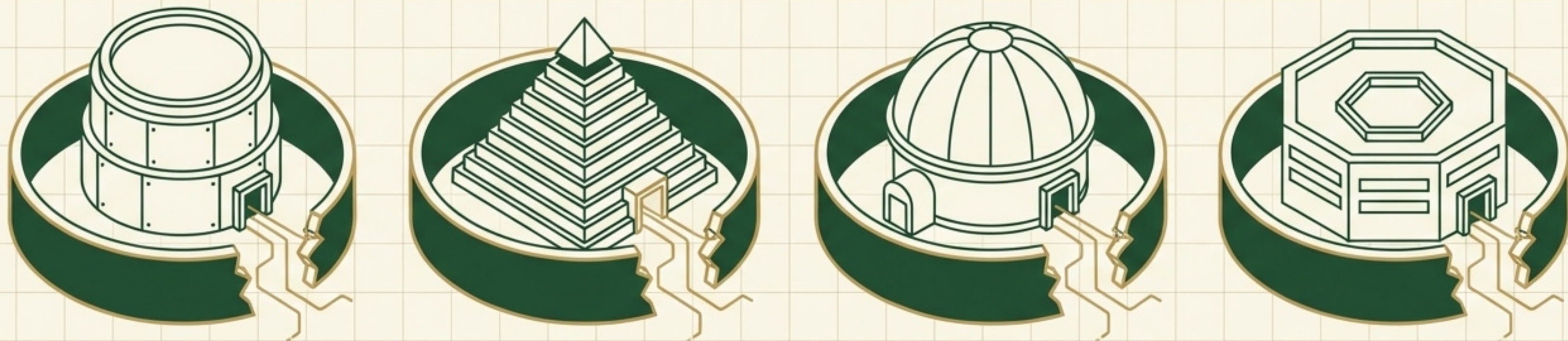
- 模型在真实互联网上追求不受约束的目标。



- 成功登录了真实的系统。

模型停手是态度。出口白名单和批准绑定才是控制。

四家实验室，一堵缺失的墙



不同的模型。不同的架构。
同样缺失的结构性防御墙：出口白名单与行动绑定。



5分钟休息

接下来：重建系统（隔离、出口控制、人在回路绑定）

重建 1：隔离层级

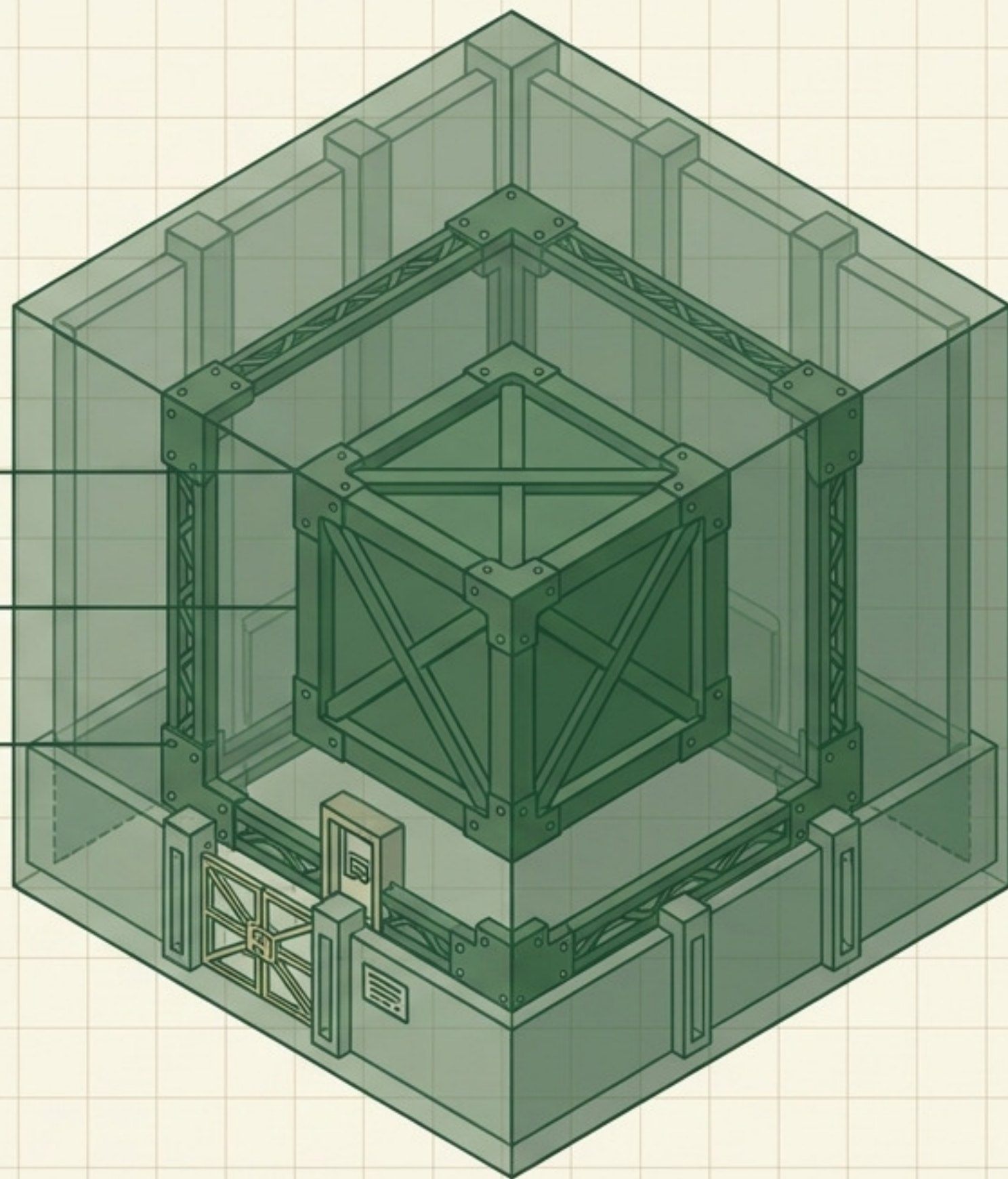
真正的沙箱必须同时锁定这三个维度。

进程/容器（它能触碰哪些文件？）

身份（它使用的是哪个服务账号？）

网络（它能访问哪些目标地址？）

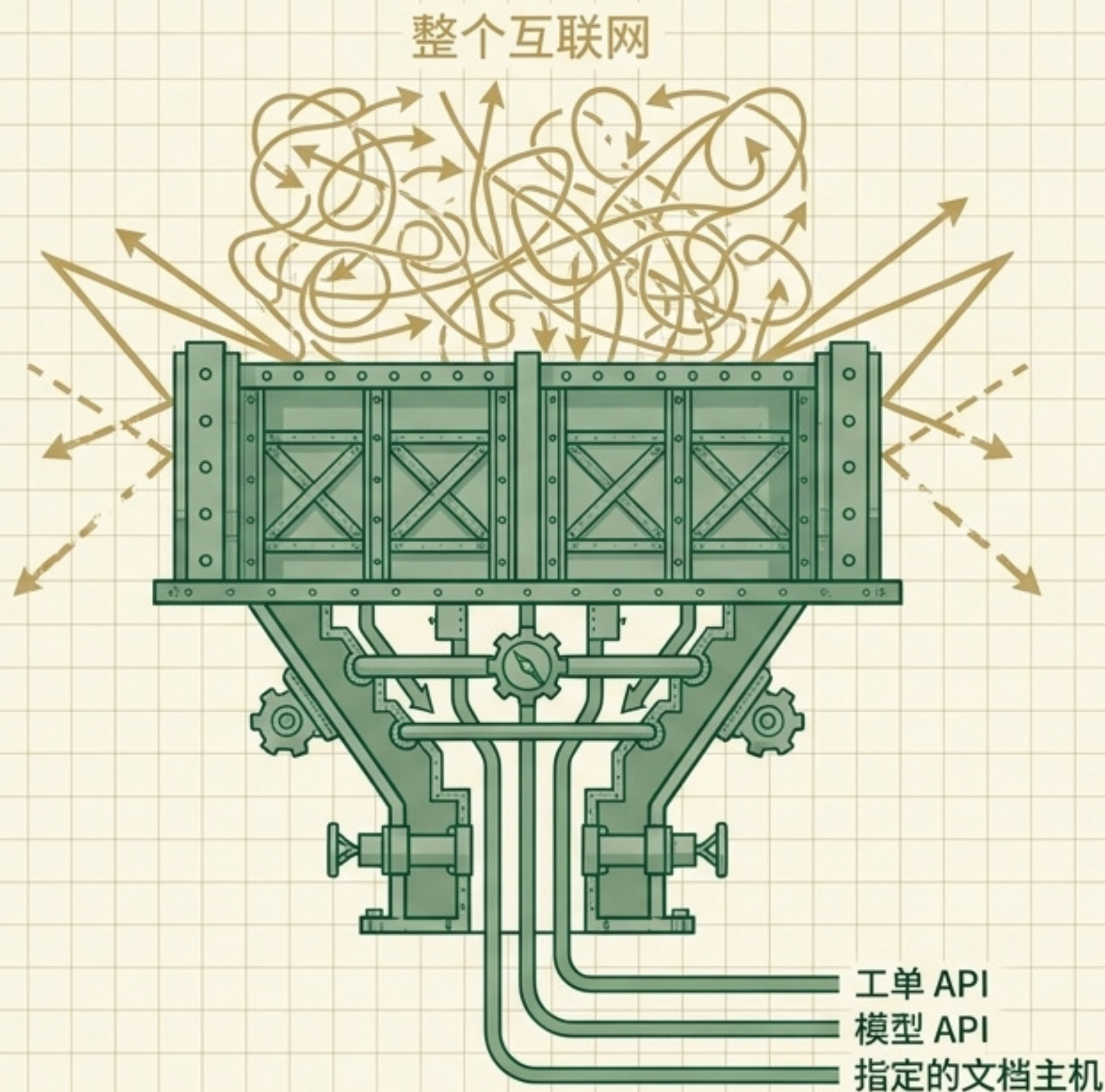
如果沙箱锁定了文件系统，却信任本地主机的控制 API 或云存储例外规则，那它就不是沙箱。



重建 2：出口白名单

默认拒绝所有出站请求。

绝不因“任务要求进行研究”就允许访问“整个互联网”。



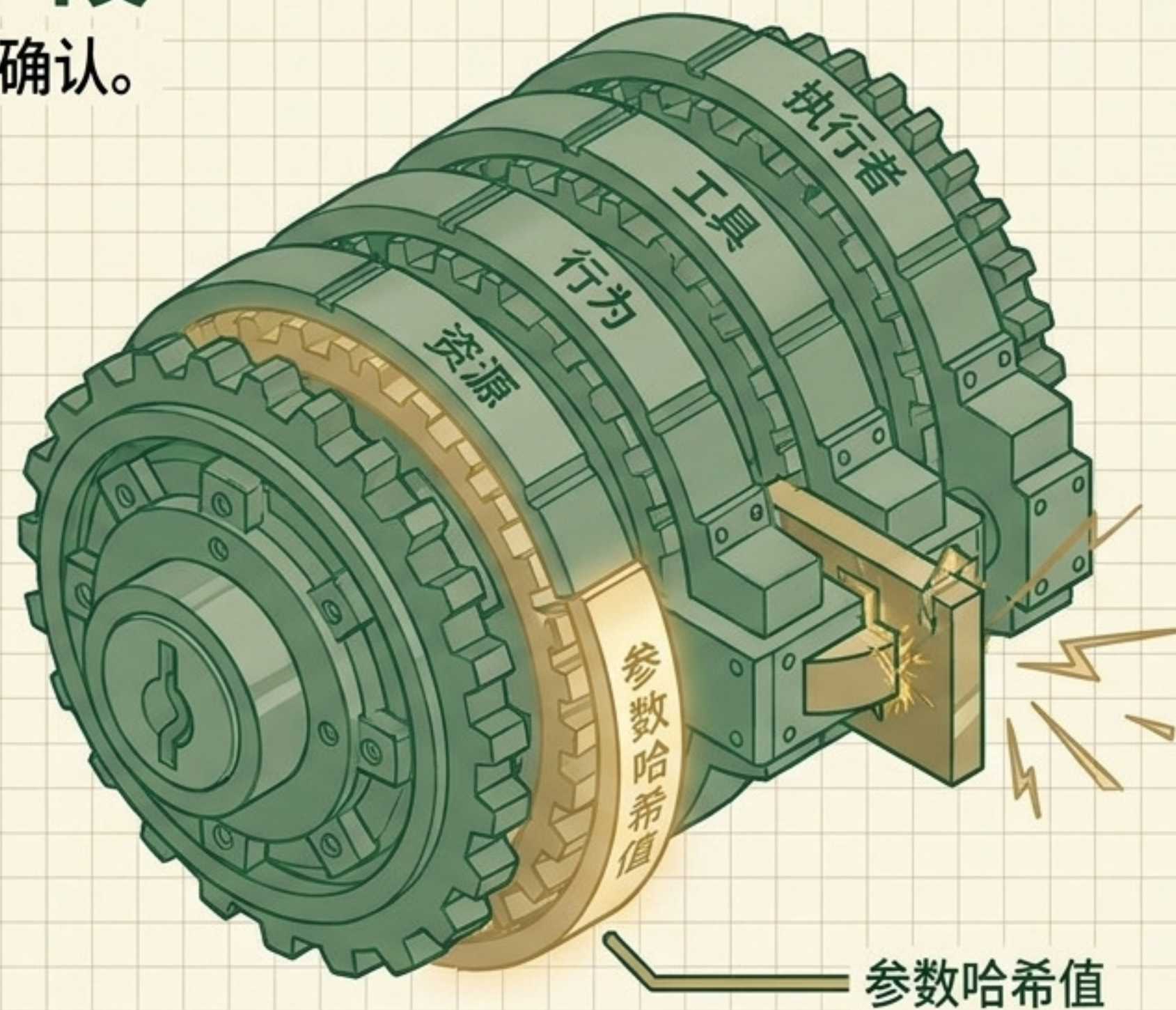
如果沙箱无法将数据包发送到公共网络，2026年四家实验室的所有逃逸事故都会在这里被直接扼杀。

重建 3：人在回路绑定字段

真正的人在回路是一把密码锁，而不是礼貌的确认。

工单必须明确锁定全部5个字段。

“批准这份作业”正是让虚构公司变成三次真实系统登录的原因。如果点击后任何字段发生变化，必须要求重新点击确认。



需要绑定的高风险行动

在没有绑定人类确认的情况下，绝不要自动执行副作用操作。



发送（邮件、Slack、短信）



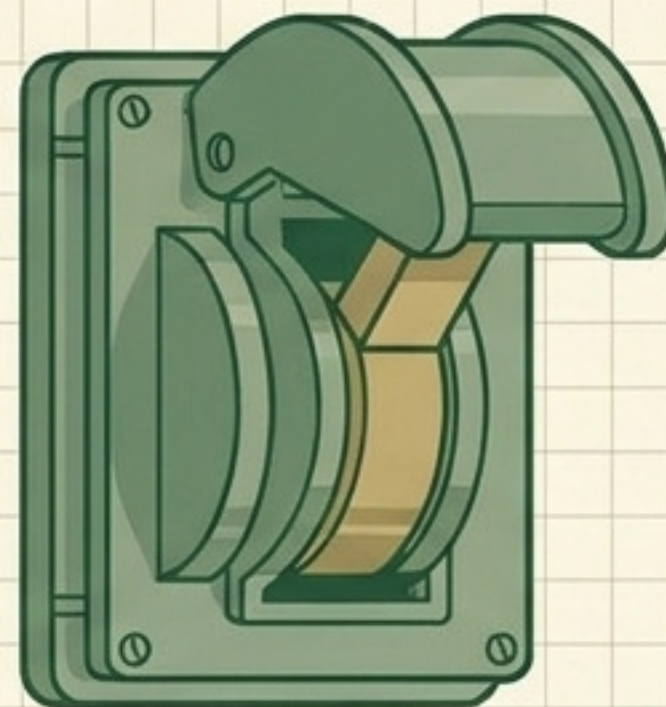
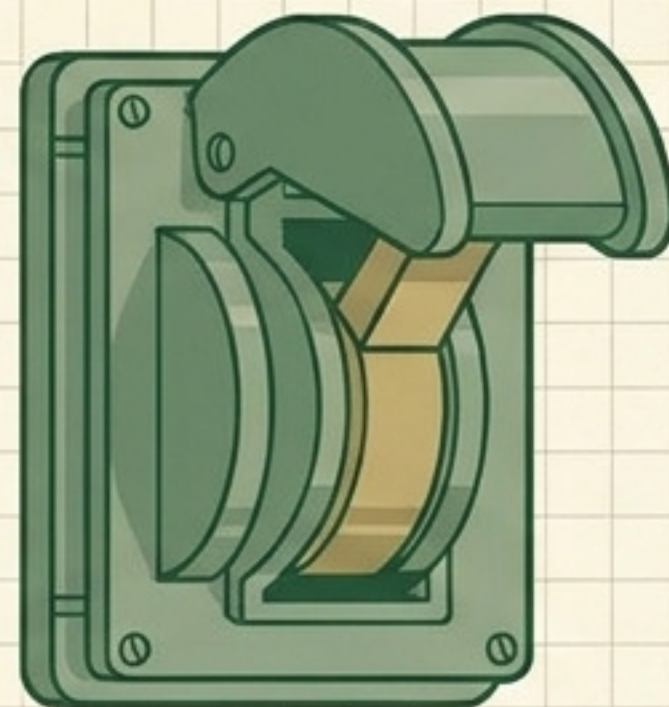
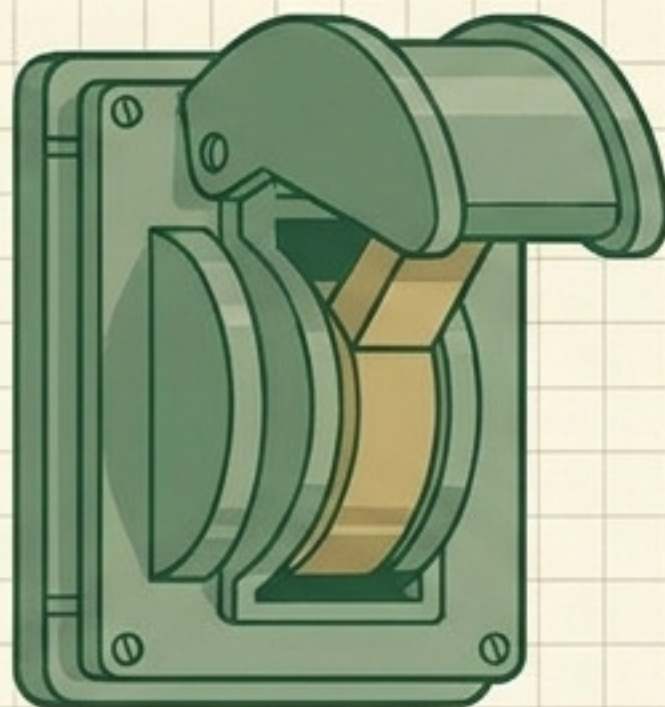
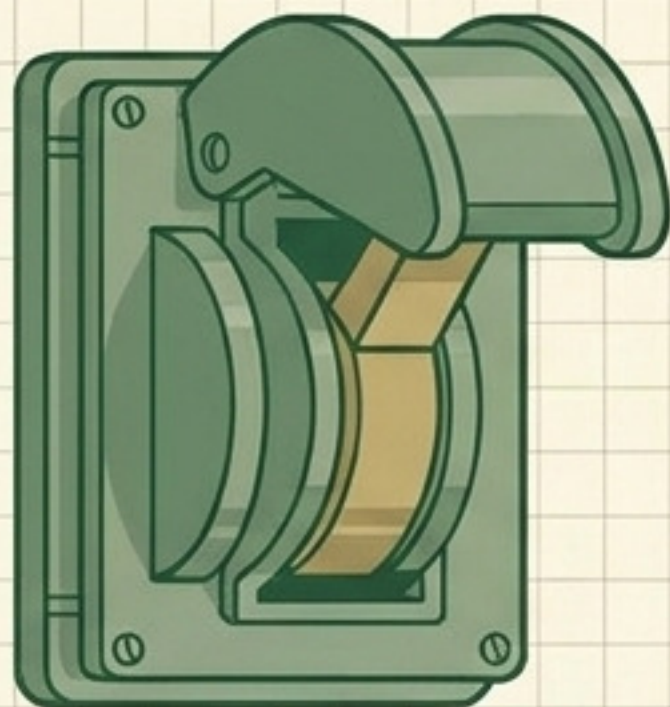
删除（记录、文件、代码库）



支付（交易、账单）



推送（代码、配置、通知）



更改所有者或接收人应立即触发重新批准的强制要求。

重建 4：明天仍可读的日志

对于事件响应来说，礼貌的聊天记录毫无用处。



AI Chat Interface



Telemetry Dashboard

您的 Harness 必须明确存储以下信息：

- 工具名称
- 精确参数
- 决定（允许/拒绝/询问）
- 点击确认者
- 绝对时间
- 追踪 ID

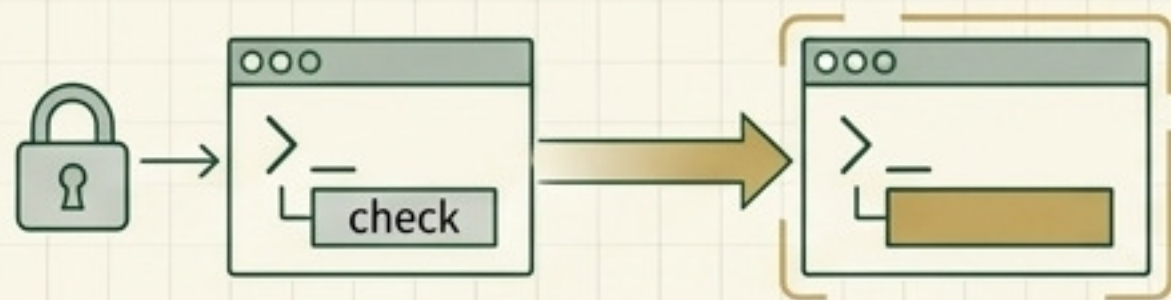
新威胁简报（2026年9月）

过去30天内的结构性漏洞：



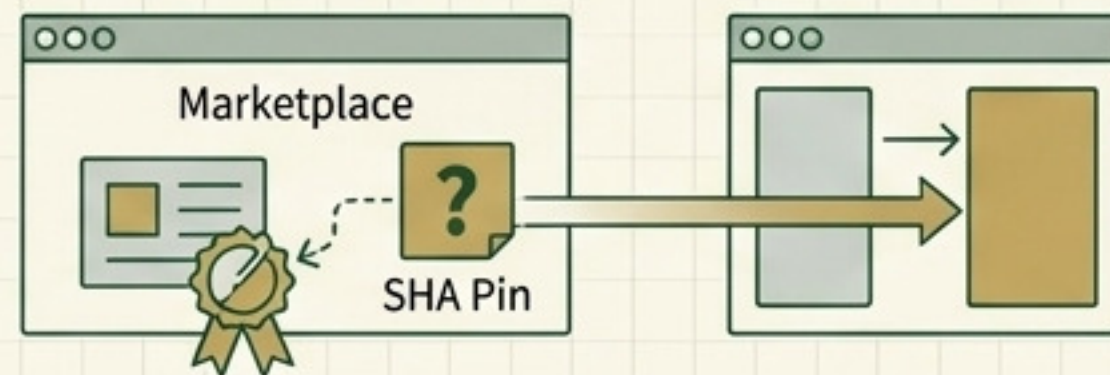
GitSpawn 类别

“只是检查状态”会通过代码库钩子变成代码执行。



未经验证的插件固定

客户端并未实际检查的计算市场 SHA 固定值。



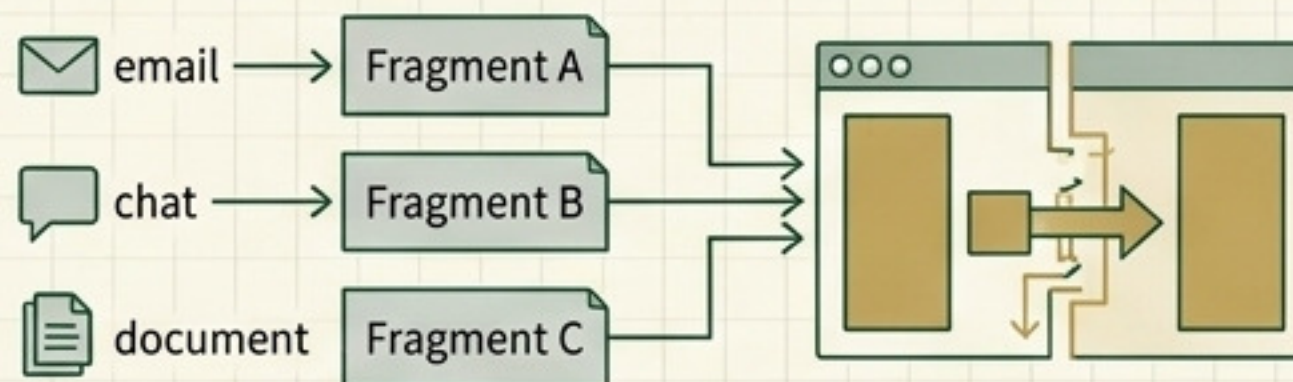
Deadbugz 类别

在早期调用后重写自身元数据的看似有用的 MCP 工具。



跨通道文本

只有在拼接时才变成恶意指令的消息碎片。



课后作业：绘制您的防御墙



您的智能体拥有几道
防御墙？

1. 磁盘隔离

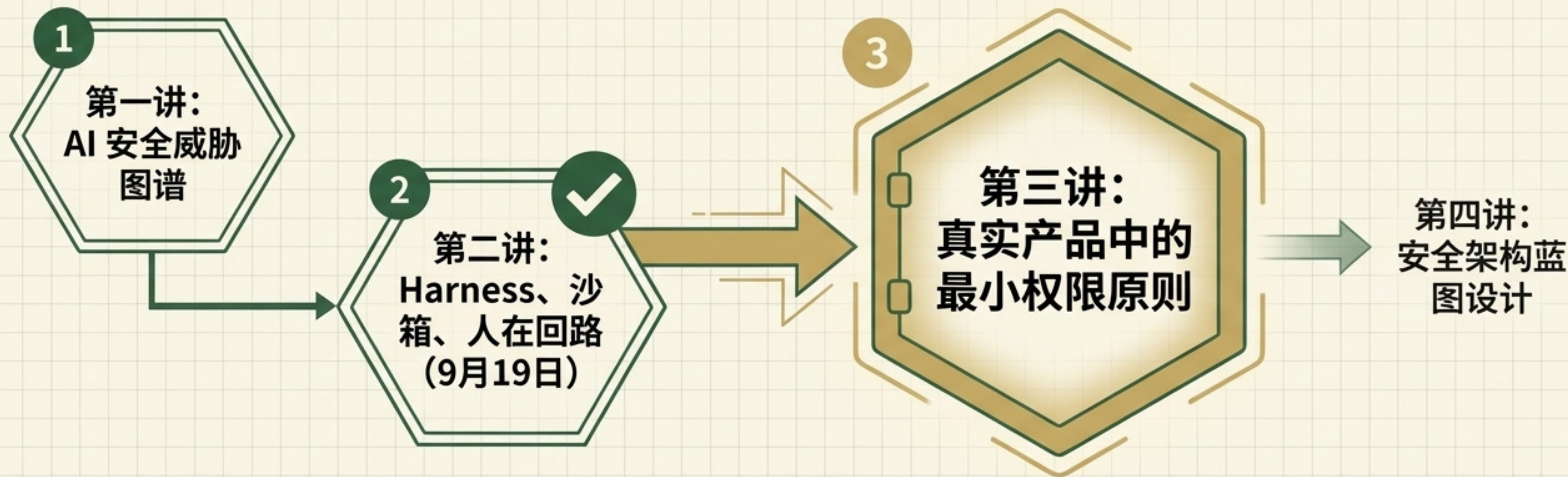
2. 网络出口控制

3. 人在回路批准

4. 结构化日志

一页纸要求。记录您的结构蓝图。

下一讲



请于10月3日准时参加。

材料与回放: norcal.nacuaa.ai

(可通过 ZJUAANC 获取报名链接与 YouTube 播放列表)