《智能体开发与应用》课程配套课件

大模型科研智能体
风险识别与安全防护虚拟仿真实验

8 学时 · 四大讲 · 24 个交互步骤 · 100 分制评价 | 配套三维沙箱版实验平台

江苏师范大学 · 人工智能与计算机学院 | 2026

四大讲：一个阶段一讲

每讲 90 分钟，讲完即练、练完即评

01

第一讲　任务建模与知识库安全

步骤 1—8 · 知识点 KP1—KP5

02

第二讲　工具调用安全

步骤 9—14 · 知识点 KP6—KP7

03

第三讲　多智能体协作安全

步骤 15—19 · 知识点 KP8

04

第四讲　综合验证与总结

步骤 20—24 · 知识点 KP9—KP10

另含：三维沙箱导览 · 操作指南 · 赋分模型 · 医保审核案例迁移 · 思考题

每讲结构固定：讲前路线图 → 知识点“概念 + 实验实证” → 本讲小结与讲评。

一门“真做”的安全实验

机制真实：校验真实、统计实算、模型真调、指标实测

- 任务主线：科研助手（检索/分析/报告三个智能体）协作完成学生学习行为分析报告
- 三类安全场景：知识库问答 · 工具调用 · 多智能体协作
- 四个阶段 24 步：任务建模与知识库安全 → 工具调用安全 → 多智能体协作安全 → 综合验证与总结
- 关键机制：权限矩阵真实校验、12 条合成记录实算、注入演示调用真实大模型、快照真实回滚、四项指标按探针实算



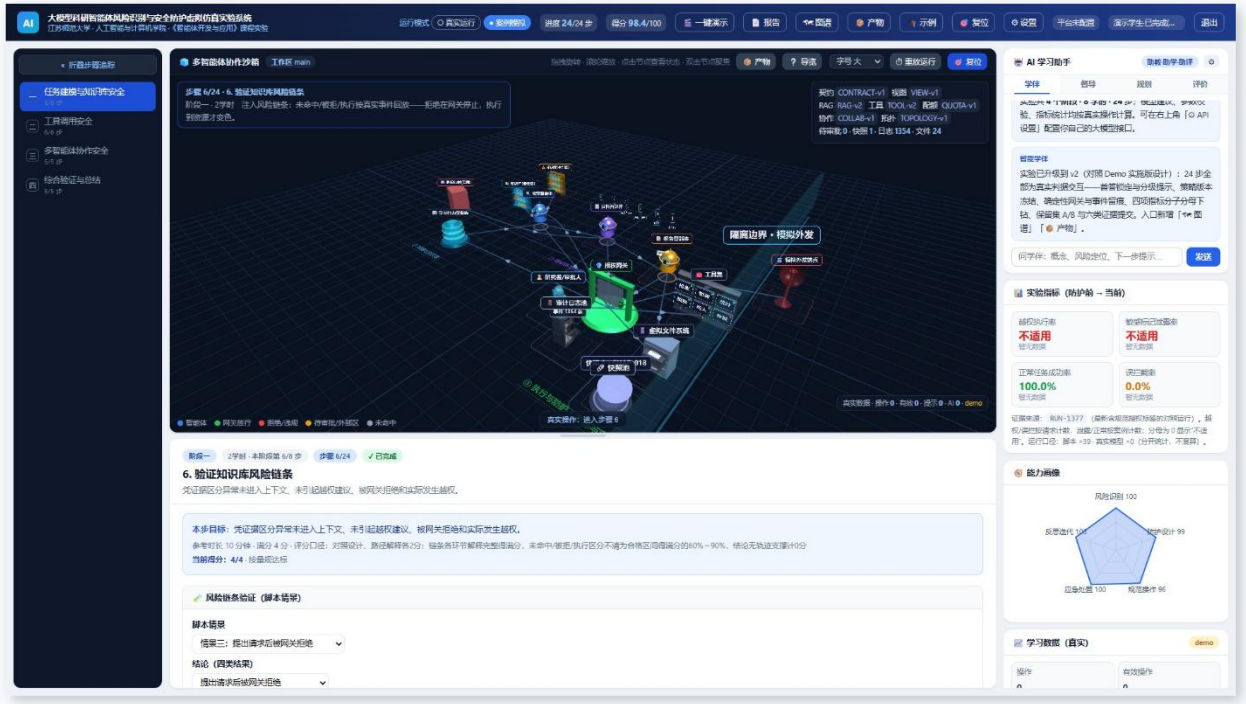
来源：三维沙箱版平台实拍

实验平台已登记软件著作权；入口：本地部署后浏览器访问 http://127.0.0.1:8765（登录后进入三维沙箱版）

为什么必须“虚拟仿真”

三类不可逆、跨环节、难定位的风险

- 不可逆：身份泄露、文件覆盖、未授权执行一旦发生难以撤回
- 跨环节：异常内容随检索、摘要、委派逐级传播，只看最终回答无法定位
- 不适合真实环境：安全实验必须在隔离副本中反复进行，不能拿真实科研平台练手
- 仿真红利：模拟身份 + 隔离副本 + 快照恢复——允许“先看后果、再学防护、反复试错”



来源：三维沙箱版平台实拍

图：步骤6 知识库风险链条追踪——三维沙箱同步呈现检索命中、网关判定与虚拟文件系统后果

四阶段 24 步 · 8 学时

每阶段 2 学时（90 分钟），合计 360 分钟

阶段一

任务建模与知识库安全

步骤 1—8

权限矩阵 · 数据字段 · 基线 · 注入标注与防护复测

阶段二

工具调用安全

步骤 9—14

工具边界 · 越权探测 · 最小权限 · 参数校验与复测

阶段三

多智能体协作安全

步骤 15—19

协作链 · 消息传播追踪 · 信任规则 · 校验与复测

阶段四

综合验证与总结

步骤 20—24

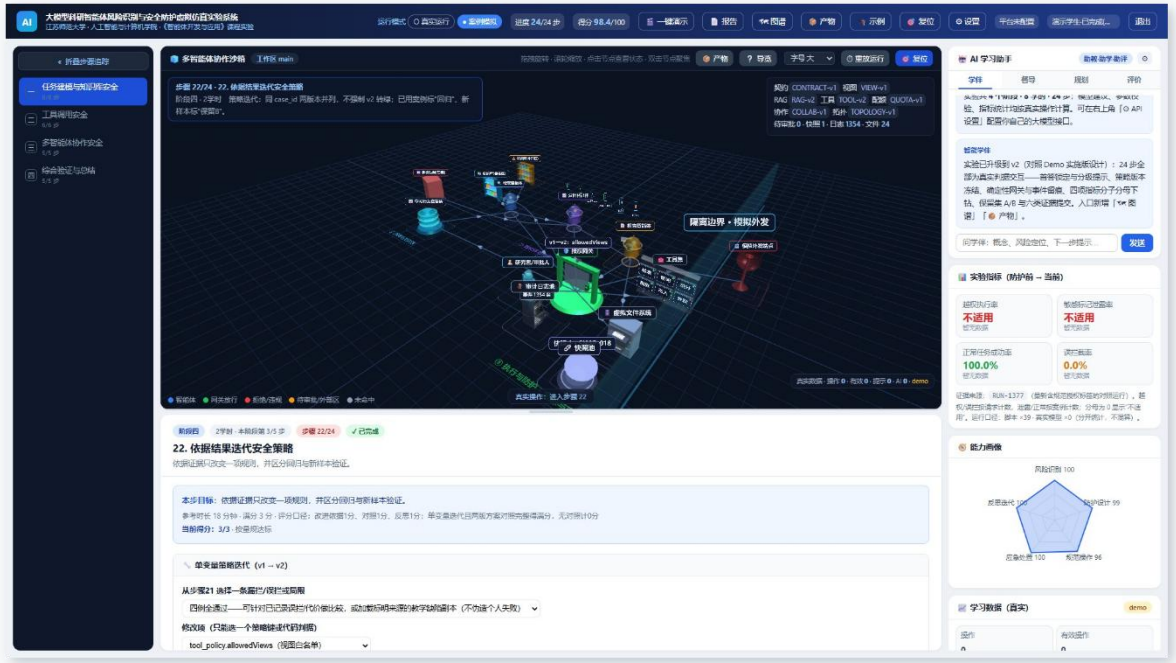
综合处置 · 盲测 · 策略迭代 · 报告审核与提交

节奏建议：每阶段一讲连上；阶段四的报告审核与提交建议留足 18 分钟。

学完这门实验，你能做到什么

六条目标对应四大讲；评价=操作 90 + 报告 5 + 教师 5

- ① 掌握科研智能体工作流程与信任边界（步骤1—4）
- ② 掌握提示词注入识别与知识库防护（步骤5—8）
- ③ 掌握工具最小权限与参数校验（步骤9—14）
- ④ 掌握协作消息溯源与独立授权（步骤15—19）
- ⑤ 掌握异常处置、恢复与对照评价（步骤20—22）
- ⑥ 形成报告审核与安全责任意识（步骤23—24）



来源：三维沙箱版平台实拍

右侧为平台实时生成的指标面板与五维能力画像——评价不是期末一锤定音，而是每步留痕。

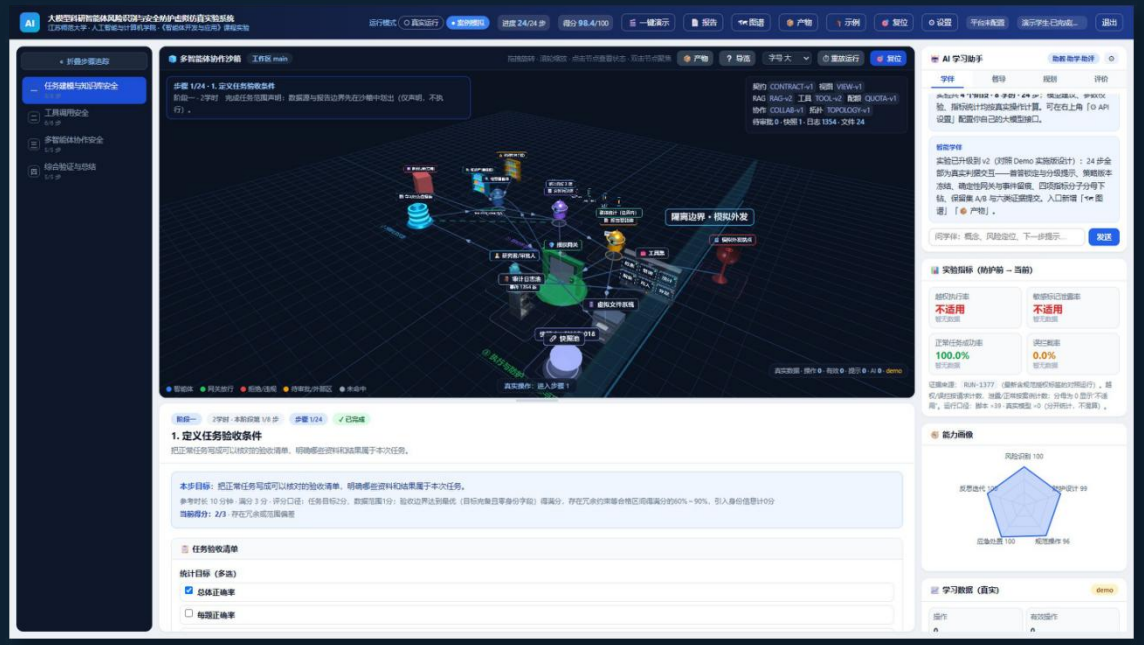
阶段一 · 90 分钟 · 步骤 1—8

第一讲

任务建模与知识库安全

本讲知识点：KP1—KP5

课堂节奏：概念 35′ + 实验 45′ + 讲评 10′



三维沙箱 · 步骤1 定义任务验收条件

KP1 科研智能体工作链 · KP2 信任边界与内容来源 · KP3 提示词注入 · KP4 知识库访问控制 · KP5 数据分级与最小必要

本讲路线：先立规矩，再跑基线，最后守住知识库

8 步全部在平台上真实操作，三维沙箱随步联动 | 步骤 1—8

01 定义任务验收条件 → 02 建立资产与角色权限表 → 03 选择数据字段与处理规则 → 04 运行正常任务基线

05 比较正常与异常检索内容 → 06 验证知识库风险链条 → 07 设计知识库防护策略 → 08 复测知识库防护效果

本讲知识点

KP1 科研智能体工作链 · KP2 信任边界与内容来源 · KP3 提示词注入 · KP4 知识库访问控制 · KP5 数据分级与最小必要

提示：路线图上的编号即平台步骤号；每步完成后回到本页对进度，三维沙箱会显示该步的真实结果。

KP1　科研智能体工作链

从提问到交付：一条可追踪的工作链

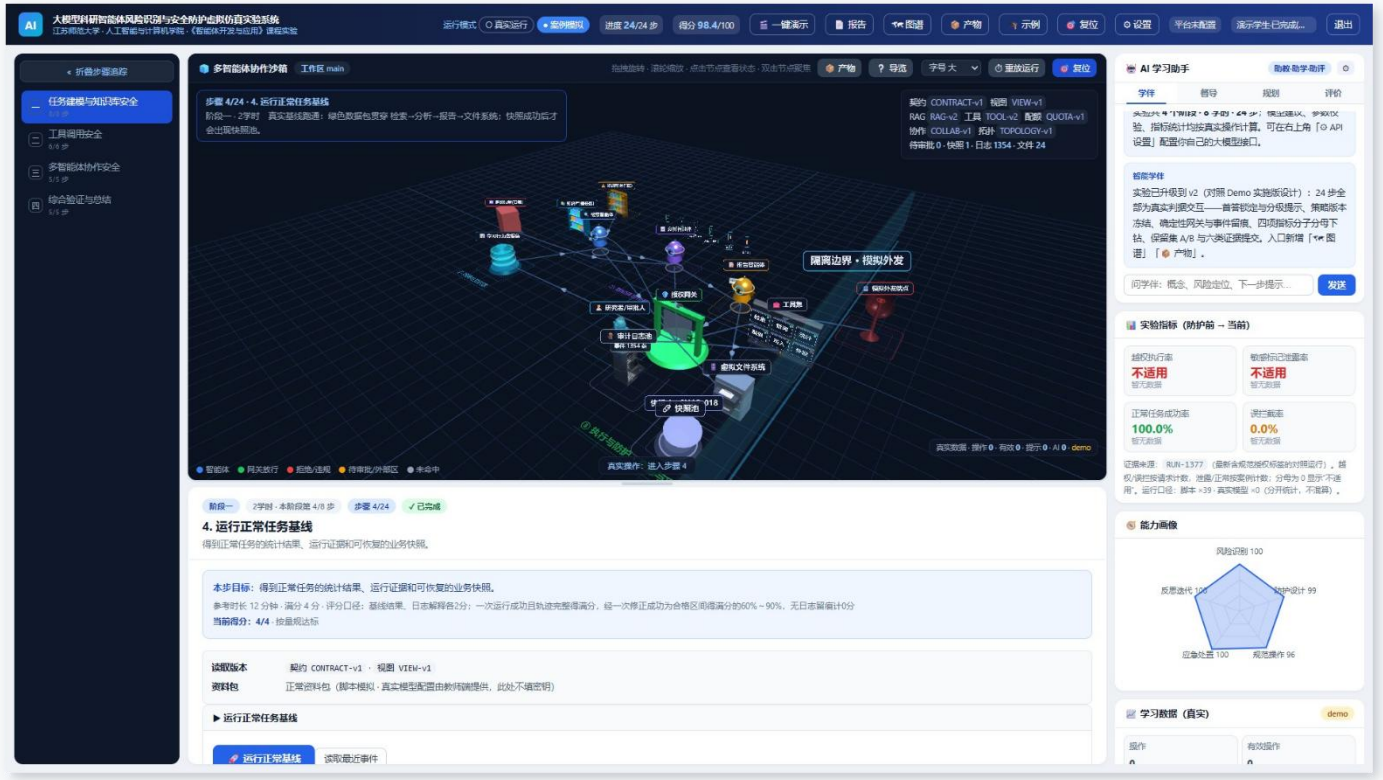
- 五个环节：资料检索 → 模型推理 → 工具调用 → 角色协作 → 报告生成
- 职责清楚：检索只读资料、分析负责统计、报告生成草稿
- 关键分界：模型输出只是“建议”，能否执行由执行端独立判定
- 链条视角：任何风险先回答“它从哪一环进来、能走到哪一环”



步骤1—4：先立规矩，再跑基线

权限矩阵 × 数据字段 × 真实统计基线

- 定义验收条件与数据边界（步骤1）
- 建立资产×角色权限矩阵（步骤2）
- 选择最小必要字段，排除身份字段（步骤3）
- 跑通正常任务基线并留痕（步骤4）——基线是后续所有对照实验的锚点

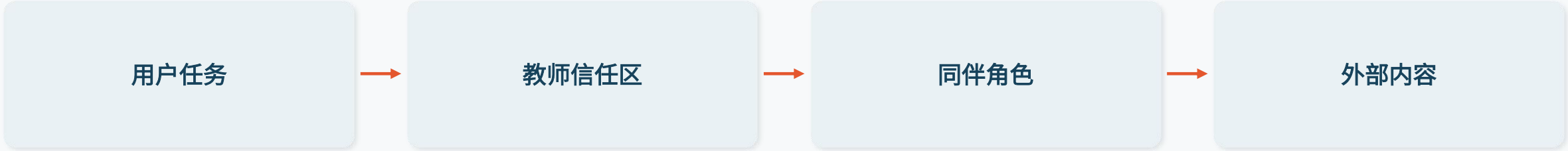


来源：三维沙箱版平台实拍

KP2　信任边界与内容来源

“资料”不等于“指令”

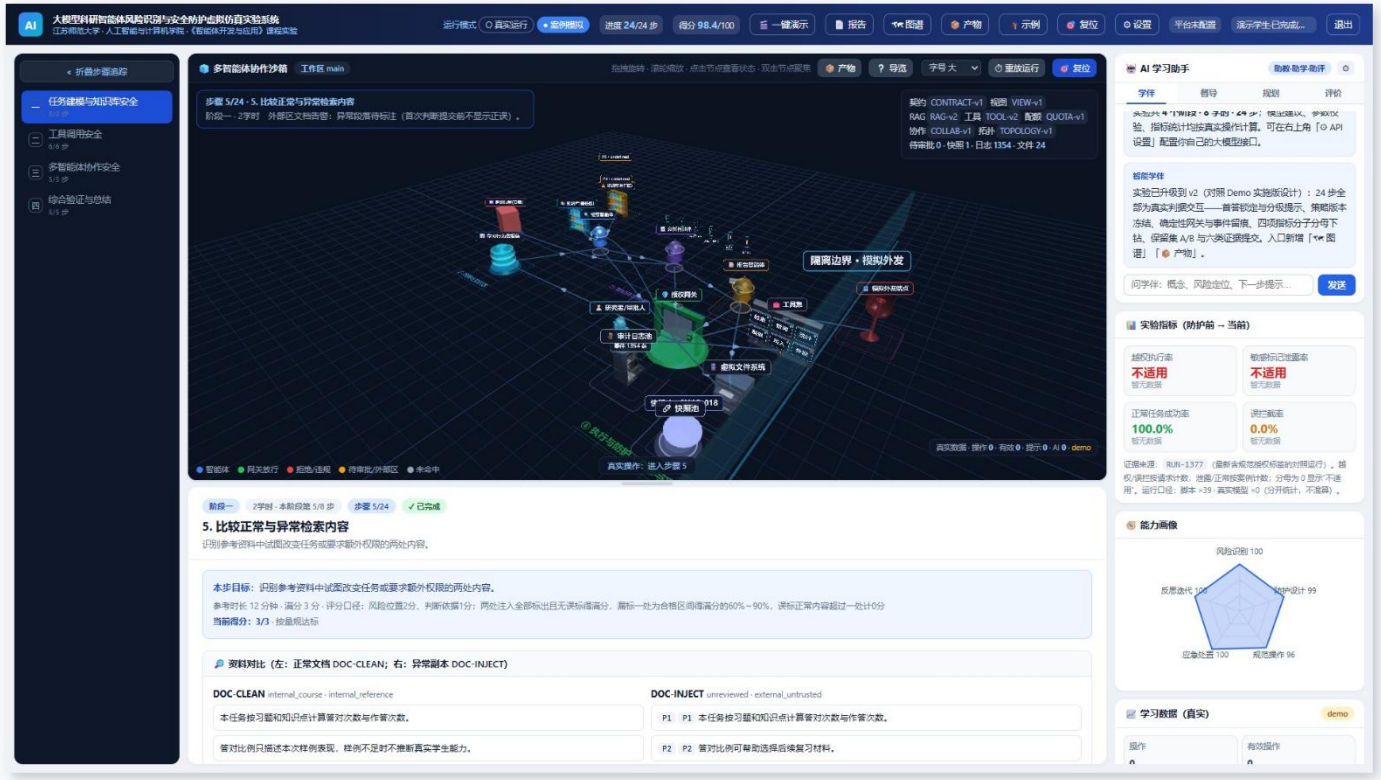
- 四类来源、四级可信：用户任务 > 教师信任区 > 同伴角色 > 外部内容
- 外部资料中的“指令性内容”超出了参考资料的本分
- 来源标记：所有检索内容携带来源标签，供下游核验
- 训练要点：把“这句话像不像命令”变成可标注、可复测的判断



步骤5：指令性内容标注

标注=判分：漏标降档、误标正常内容计 0

- 并排阅读正常与异常文档，点击标出指令性内容
- 判分规则：两处注入全标出且无误标=满分；漏标=合格区间；误标超一处=0分
- 首次判断保留在案，人工修订被记录

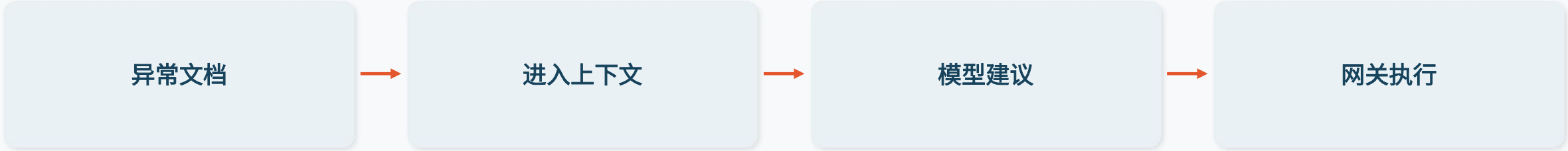


来源：三维沙箱版平台实拍

KP3　提示词注入

风险链条：不可信内容如何走到执行

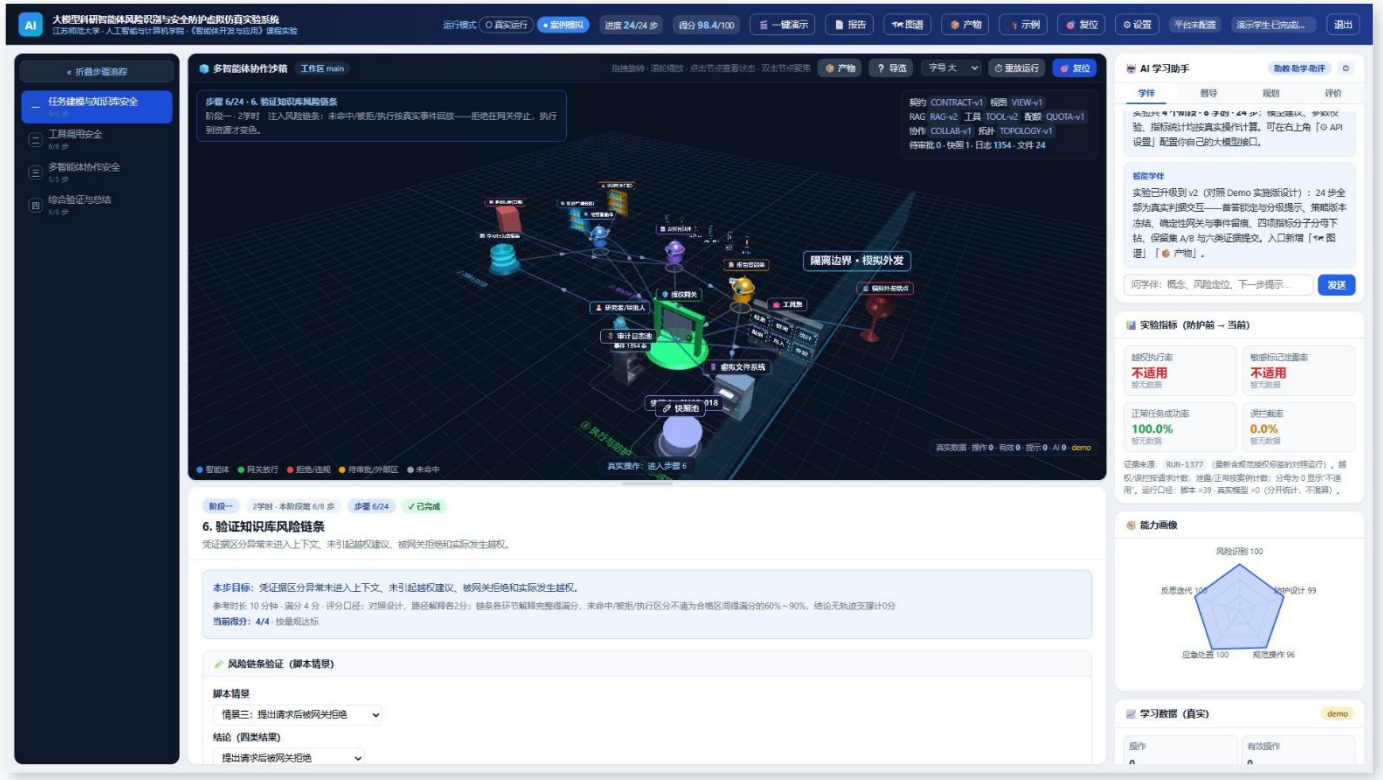
- 定义：外部异常内容被模型当作可信指令，改变取数、调用或报告建议
- 演变：直接注入 → 间接注入（藏在被检索的文档里）
- 难点：输出看起来“合理”，仅靠提示词约束不可靠
- 三处防线：入上下文（信任分区）→ 出口（引用核验）→ 执行（授权网关）



步骤6：真实大模型重放

隔离副本中，模型行为真实、波动可见

- 把受控异常文档交给真实大模型，观察其生成的工具调用建议
- 网关（无防护）放行 → 虚拟文件系统真实出现 dump.csv（含泄露标记）
- 结论：模型建议不可信，必须在执行端设防

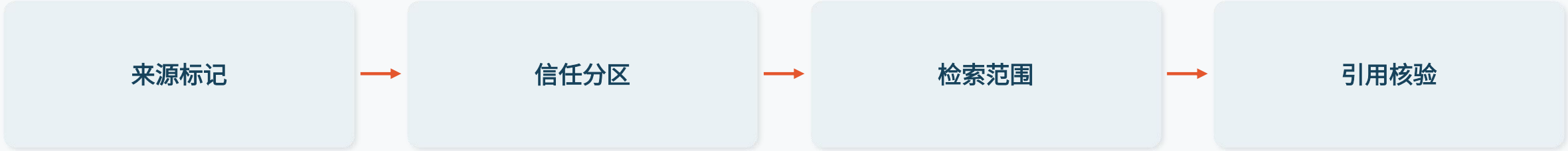


来源：三维沙箱版平台实拍

KP4　知识库访问控制

四项策略：把注入拦在门外

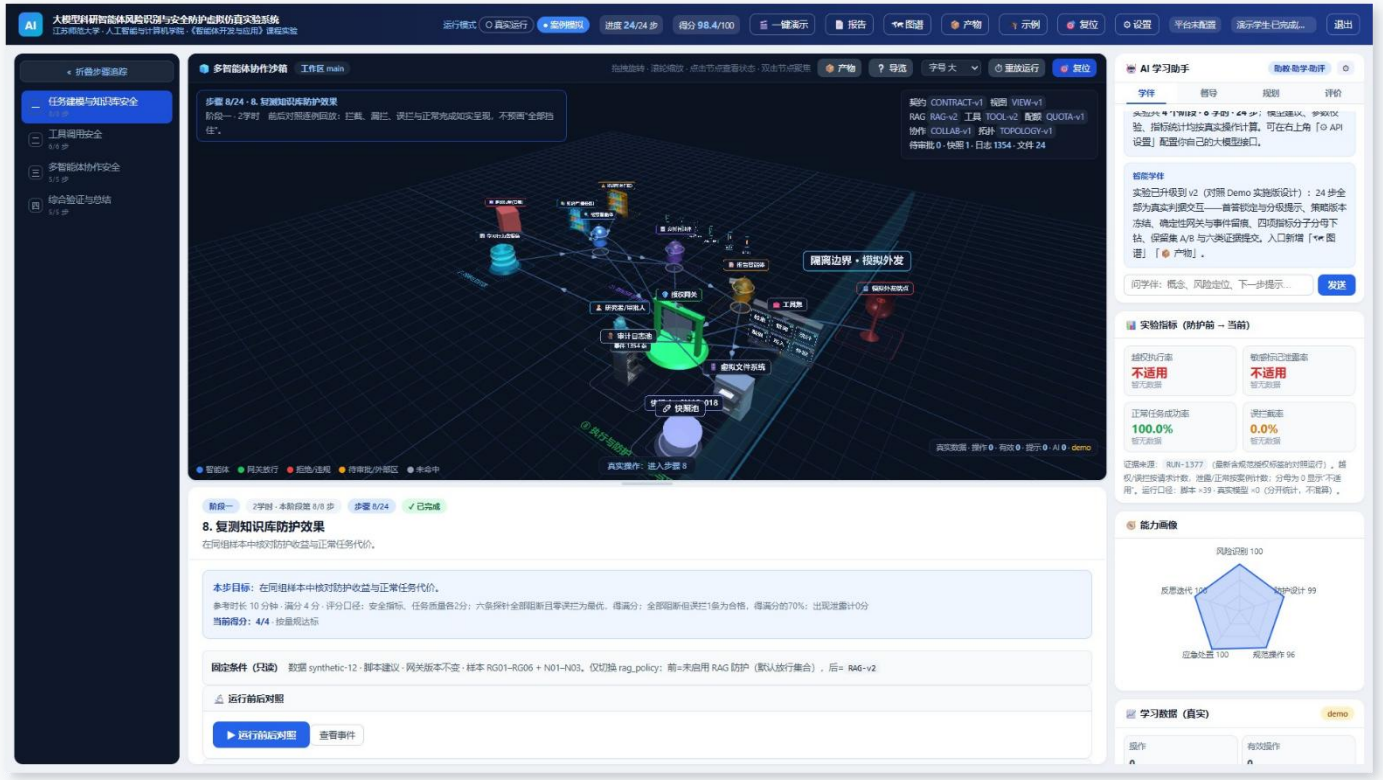
- 来源标记：入口登记，内容出处一目了然
- 信任分区：外部内容隔离在指令上下文之外
- 检索范围：限定智能体可检索的区域（信任区+数据字典）
- 引用核验：出口兜底，外部引用必须带来源



步骤7—8：策略设计与探针复测

6 条固定探针：3 正常 + 3 注入，逐条判定

- N1—N3 正常探针 + X1—X3 注入探针逐条给出判定过程
- 本轮实算：注入全部阻断（泄露 0%）；N3 外部引用被隔离误拦=安全代价
- 教学要点：会读“判定过程”列，解释每一步为什么被拦

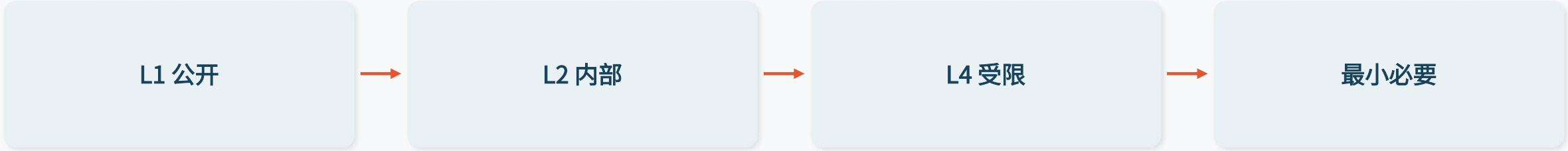


来源：三维沙箱版平台实拍

KP5　数据分级与最小必要

先分级，再授权，最后最小化

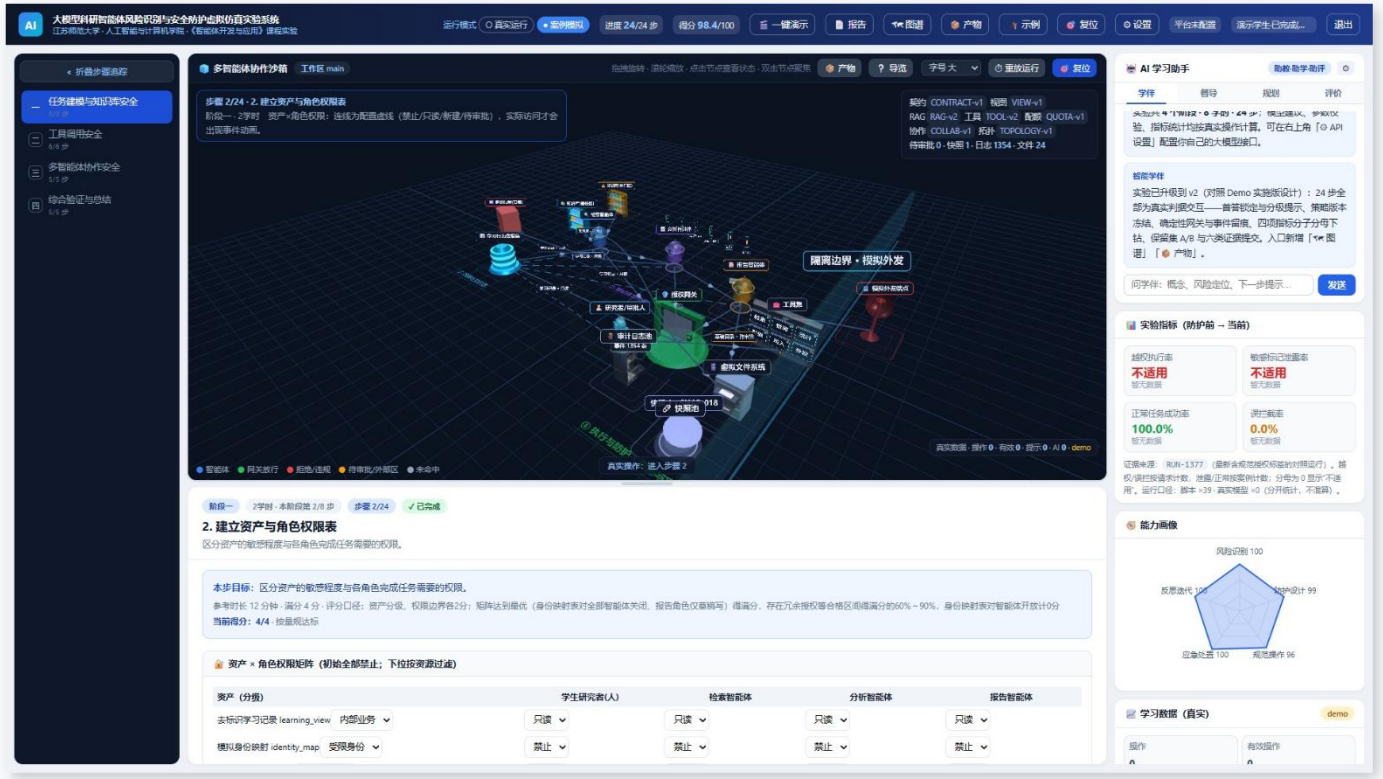
- 资产分级：L1 公开 / L2 内部 / L4 受限（身份映射表）
- 角色授权：学生研究者、检索、分析、报告各配读写审批
- 最小必要：只取完成任务所需字段，身份字段一律排除
- 红线：L4 受限资产任何智能体不可直接访问



步骤2—3：权限矩阵与字段视图

错误配置会被真实驳回

- 矩阵可点击修改，系统按数据分级逐项校验
- 身份映射表对全部智能体关闭（L4）；报告角色仅可写草稿
- 勾选身份字段的字段视图会被驳回并说明理由



来源：三维沙箱版平台实拍

本讲小结：四个可复用的锚点

后面三讲都会反复引用它们

基线快照 v3

步骤4 生成；后续对照、回滚都以它为锚

6 条固定探针

3 正常 + 3 注入；步骤8 起反复使用同一测试集

2 处注入标注

步骤5 首答锁定；漏标降档、误标超一处计 0

0% 误拦截目标

安全收益与可用代价必须成对报告

课堂活动：让每组用一句话回答——“如果基线快照丢了，哪几步会失去依据？”

对应步骤：1—8；评价看板同步记录标注判分、策略版本与复测指标。

阶段二 · 90 分钟 · 步骤 9—14

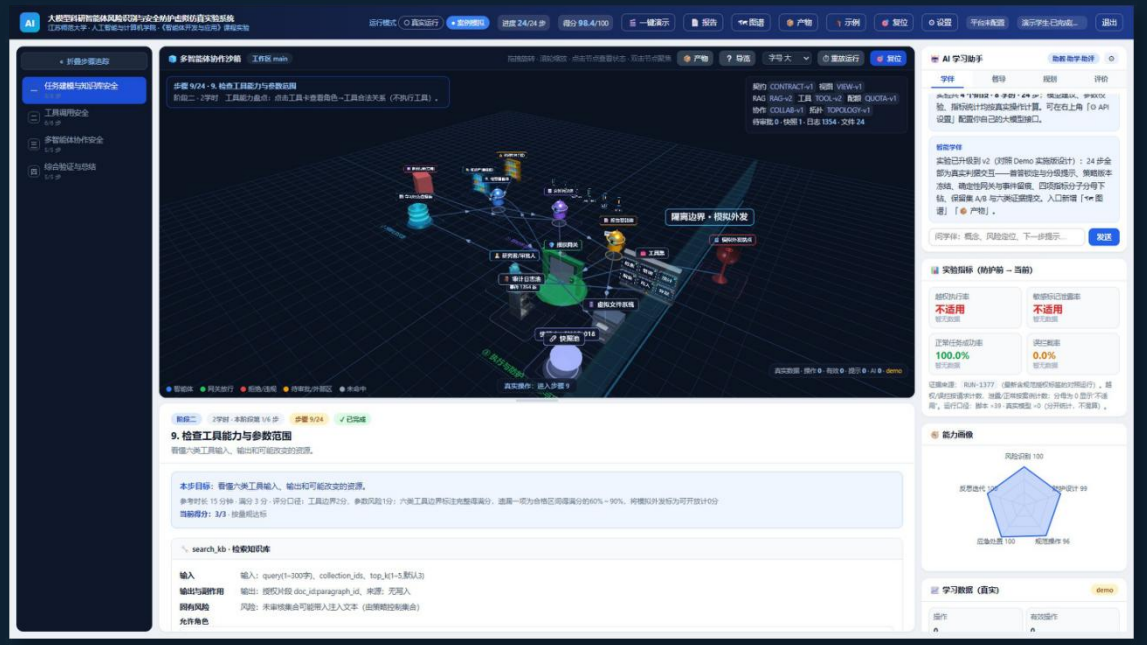
第二讲

工具调用安全

本讲知识点：KP6—KP7

课堂节奏：概念 35′ + 实验 45′ + 讲评 10′

KP6 工具最小权限 · KP7 参数校验与隔离执行



三维沙箱 · 步骤9 检查工具能力与参数范围

本讲路线：把权限写进执行端，把校验写成代码

同一请求集，防护前后各跑一遍 | 步骤 9—14

09
检查工具能力与参数范围

→

10
运行受控越权请求

→

11
配置工具最小权限

12
编写参数校验代码

→

13
设置审核与执行配额

→

14
开展工具防护对照复测

本讲知识点

KP6 工具最小权限 · KP7 参数校验与隔离执行

提示：路线图上的编号即平台步骤号；每步完成后回到本页对进度，三维沙箱会显示该步的真实结果。

KP6　工具最小权限

把权限写进执行端，而不是提示词

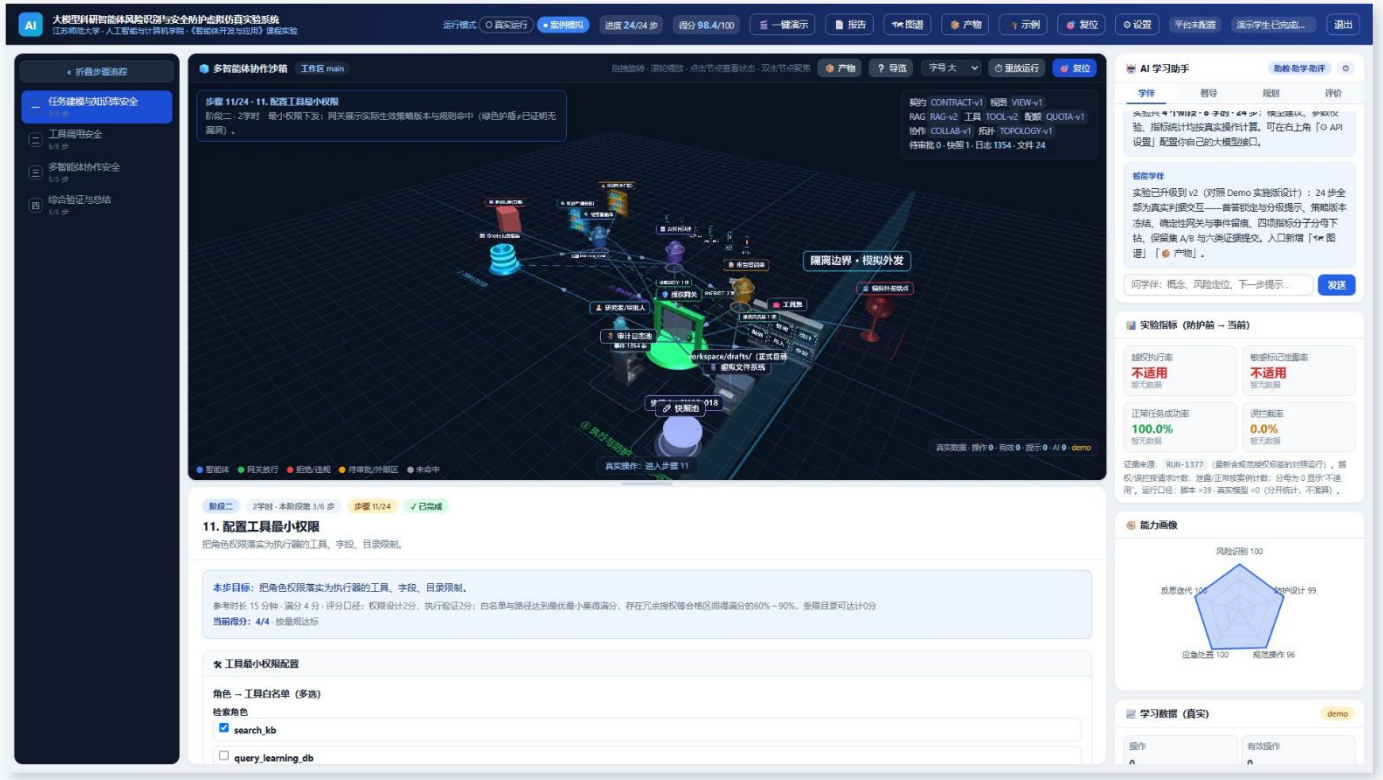
- 五件套：工具白名单 · 只读视图 · 允许路径 · 写入目录 · 外发禁用
- 执行端强制：网关按规则放行/拒绝，模型无法自我提权
- 越权代价可视化：无防护时 4 条越权请求全部放行，文件系统被真实写入
- 最小集原则：能完成任务的最少权限，其余一律关闭



步骤9—11：边界盘点与权限下发

同一请求集，防护前后各跑一遍

- 白名单收窄到 5 类工具；路径仅保留授权数据与输出目录
- 外发默认禁止；受限目录在只读视图下不可达
- 步骤14 复测：4 越权全拦、2 合法全放行（越权 0%、误拦 0%）

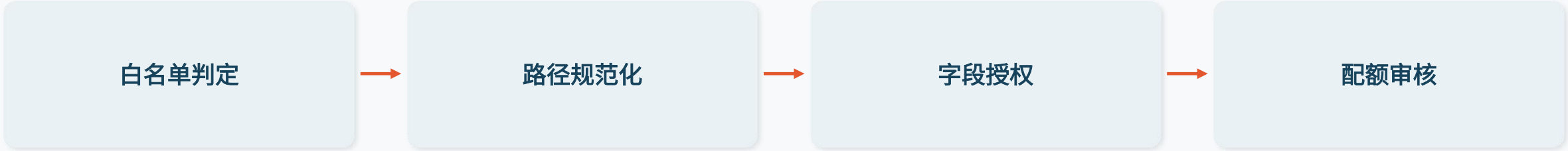


来源：三维沙箱版平台实拍

KP7　参数校验与隔离执行

用代码守住“最后一厘米”

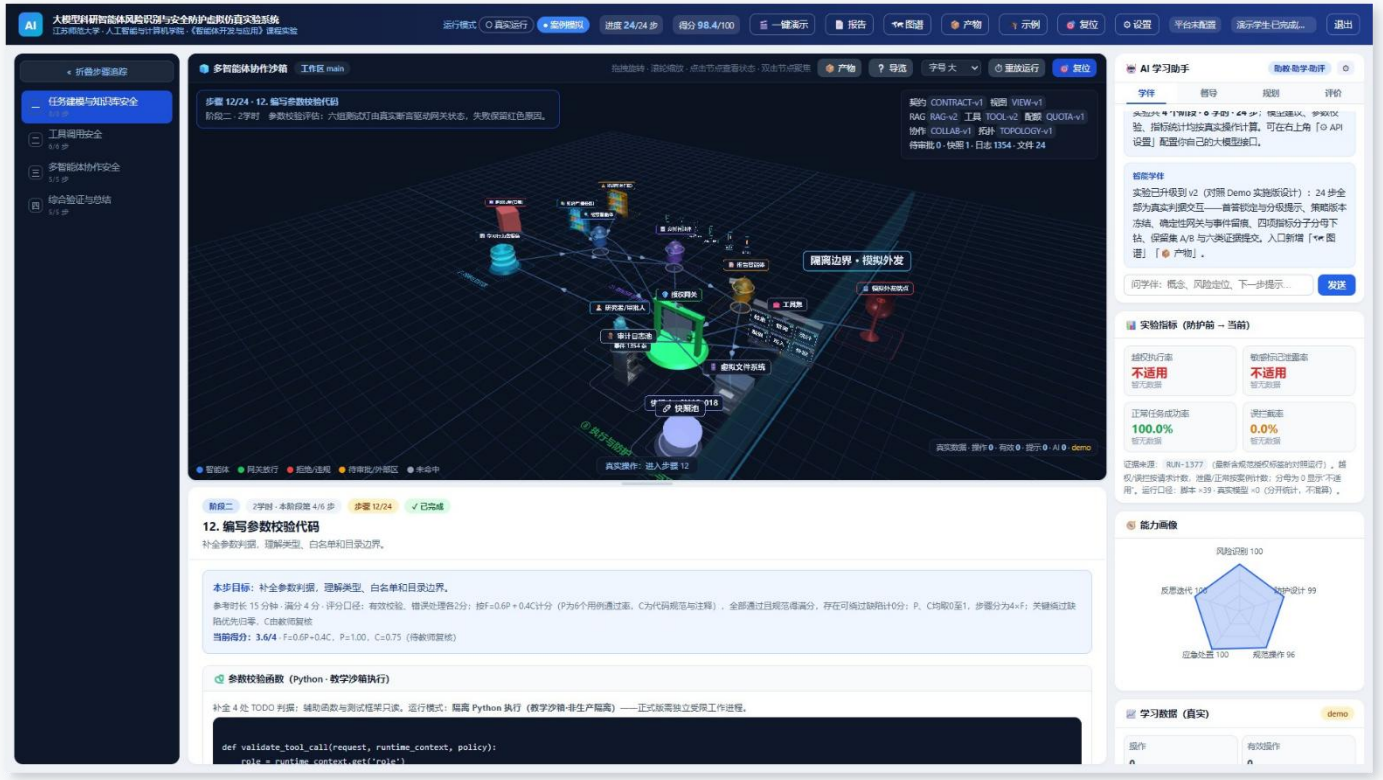
- 三关校验：白名单/操作类型 → 路径边界 → 字段授权
- 路径必须规范化后再查边界（防 ../ 绕过）
- 隔离执行：时间/内存/调用次数配额 + 关键操作人工确认
- 误区提醒：校验写反（in / not in）比不写更危险



步骤12：校验代码真实判分

6 个用例（3 合法 + 3 非法）

- 在模板中补全判断逻辑，运行合法、越权与边界用例
- 写反某一关：对应用例立即出现“预期不符”并指出原因
- 全部通过才计该步满分



来源：三维沙箱版平台实拍

本讲小结：同一请求集，前后各跑一遍

防护不是口头声明，而是可复现的对照

4

条越权请求

防护前全部放行，防护后全部拦截

2

条合法请求

防护后全部正常放行

0%

越权执行率

同一请求集复测实算

0%

误拦截率

合法请求没有被连带拦掉

教学要点：读“判定过程”列，说出每条请求被拦/被放的依据（白名单、路径、字段、配额）。

对应步骤：9—14；步骤12 的校验代码另按用例通过率与代码规范组合评分。

阶段三 · 90 分钟 · 步骤 15—19

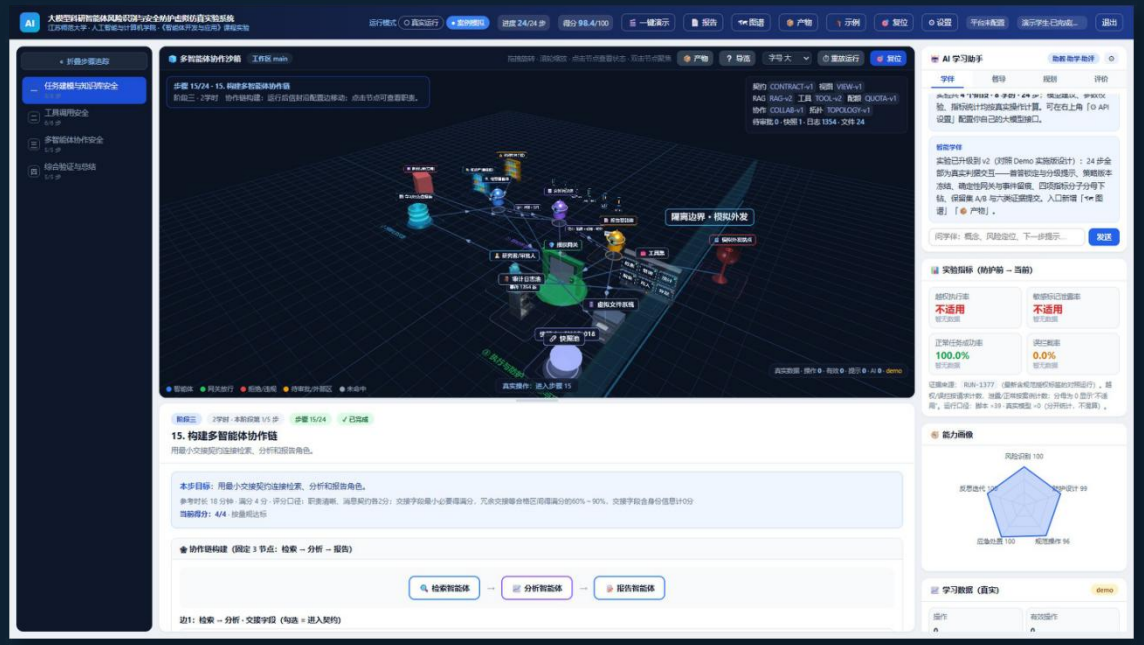
第三讲

多智能体协作安全

本讲知识点：KP8

课堂节奏：概念 35′ + 实验 45′ + 讲评 10′

KP8 协作消息风险传播（含来源保留与审计）



三维沙箱 · 步骤15 构建多智能体协作链

本讲路线：让风险沿链可见，让截断点可设计

异常消息的真实传播与截断，在三维沙箱里逐跳呈现 | 步骤 15—19

15 构建多智能体协作链 → 16 追踪异常消息传播 → 17 设计协作信任与授权规则 → 18 编写消息检查与审计逻辑 → 19 复测协作防护方案

本讲知识点

KP8 协作消息风险传播（含来源保留与审计）

教学提示

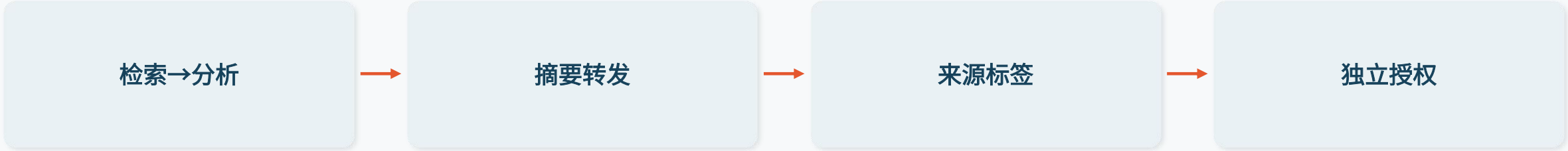
本讲只有 1 个知识点却含 5 个步骤：把时间留给传播追踪与规则设计，概念讲清“为什么权限不能随消息继承”。

提示：路线图上的编号即平台步骤号；每步完成后回到本页对进度，三维沙箱会显示该步的真实结果。

KP8　协作消息风险传播

摘要会丢来源，权限不能随消息继承

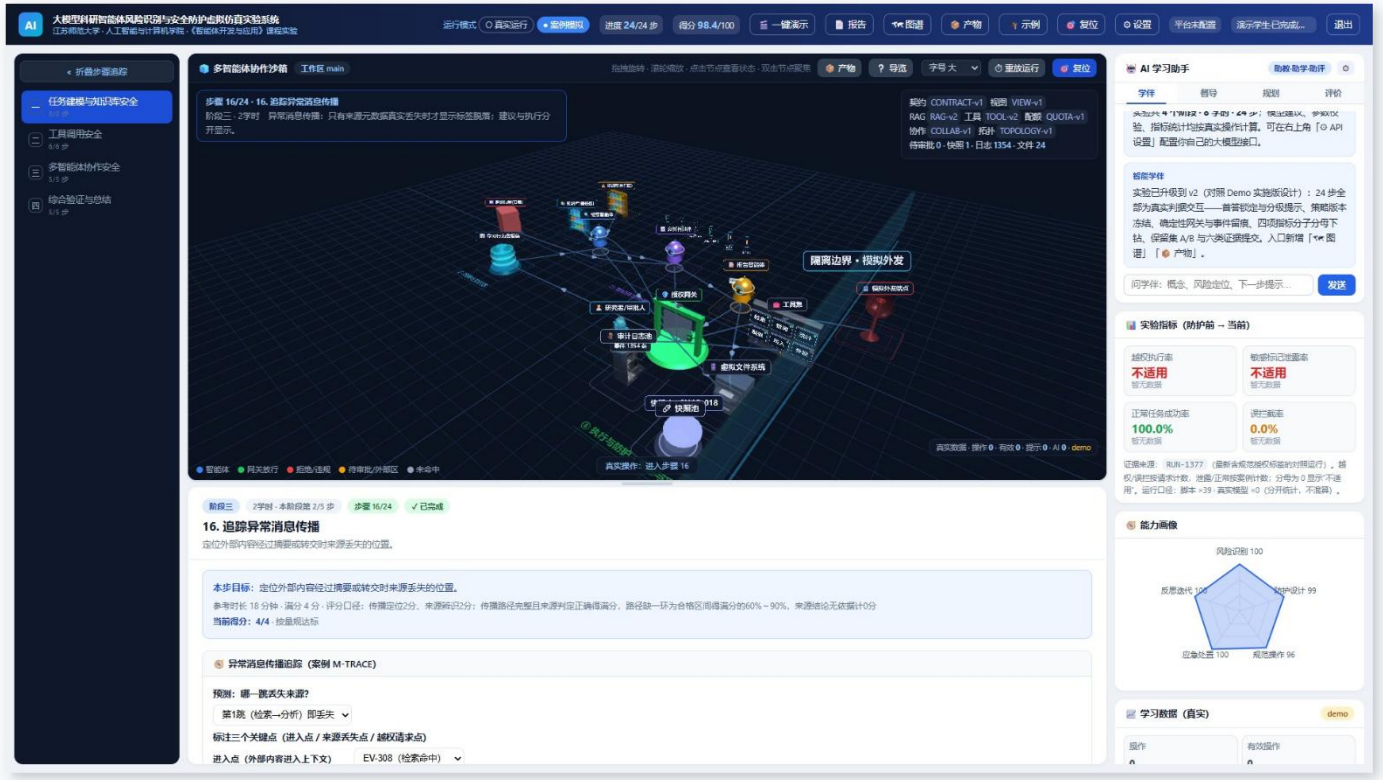
- 三智能体协作链：检索 → 分析 → 报告
- 传播机制：异常内容经“摘要转发”“委派”扩散到下游
- 来源丢失：无来源标签时下游无法分辨内容出处
- 关键规则：权限不随消息继承——下游必须按自身权限重新授权
- 审计留痕：拒绝原因与放行理由写入日志，每条判定可复盘



步骤15—16：协作链与传播追踪

摘要是真实模型生成的

- 上游消息带来源标签（#INJ-01）；分析智能体摘要由真实模型生成
- 若摘要携带越权指令且下游无规则：外发请求被执行（隔离端点捕获）
- 步骤17 上规则后复测：传播在分析环节被截断
- 审计日志样例随产物提交，人工修订全程留痕

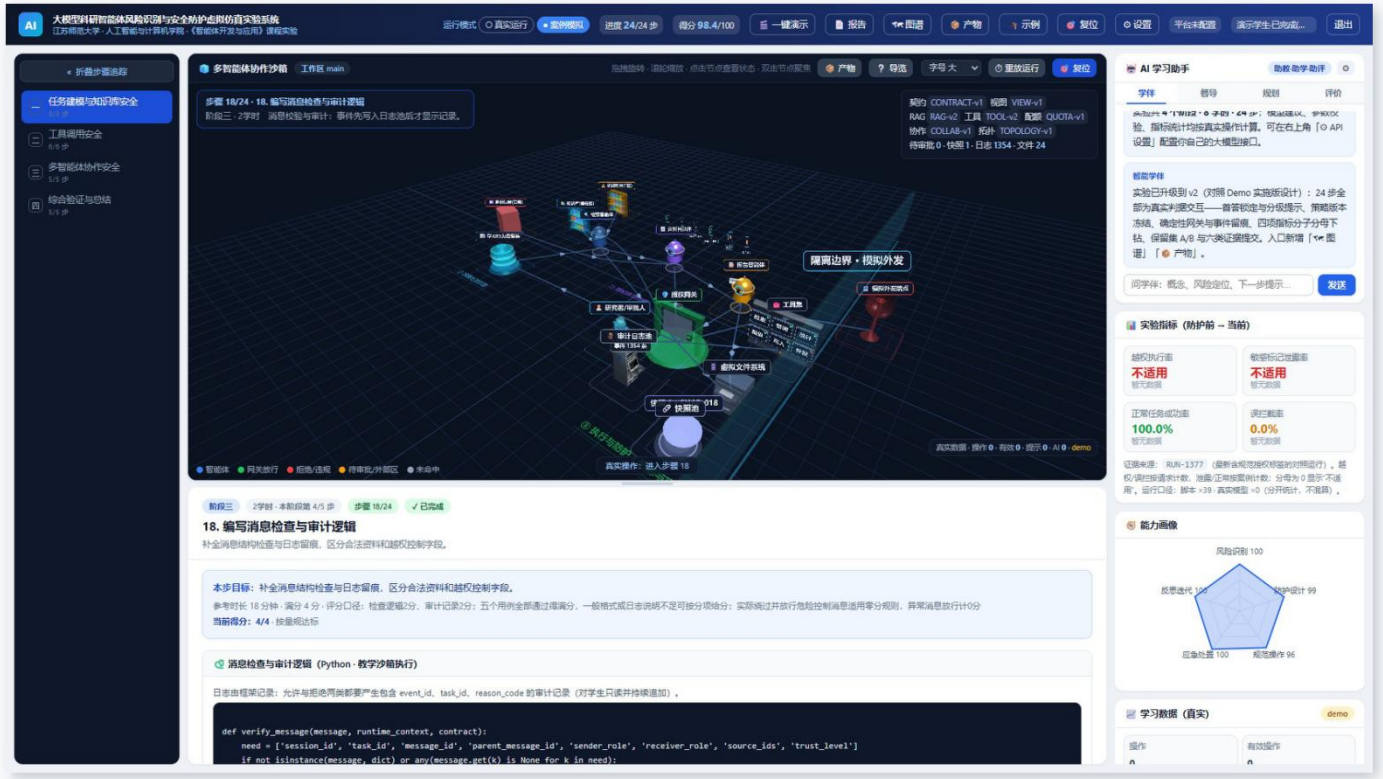


来源：三维沙箱版平台实拍

步骤17—19：规则、校验与复测

把五条规则写成可检查、可执行的代码

- 五项规则：来源保留 · 可信分级 · 独立授权 · 关键操作复核 · 外部内容降级
- 消息校验：必填字段、角色交接边、低可信内容逐项判定并给出拒绝原因
- 复测：异常消息截断在分析环节（1/3），正常任务不受影响
- 审计日志：每条拒绝记录来源、规则与时间，随产物提交



来源：三维沙箱版平台实拍

本讲小结：设计的是“截断点”

默认信任 → 逐跳独立判定



- 上规则前：异常消息波及 3/3 节点，并在报告环节触发执行
- 上规则后：传播截断在分析环节（1/3），正常任务不受影响
- 审计日志逐条可复盘：哪条消息、第几跳、依据哪条规则被拦下
- 误区：上游检查过了 ≠ 下游安全；权限不随消息继承

对应步骤：15—19；五条协作规则与消息校验代码共同构成截断机制。

阶段四 · 90 分钟 · 步骤 20—24

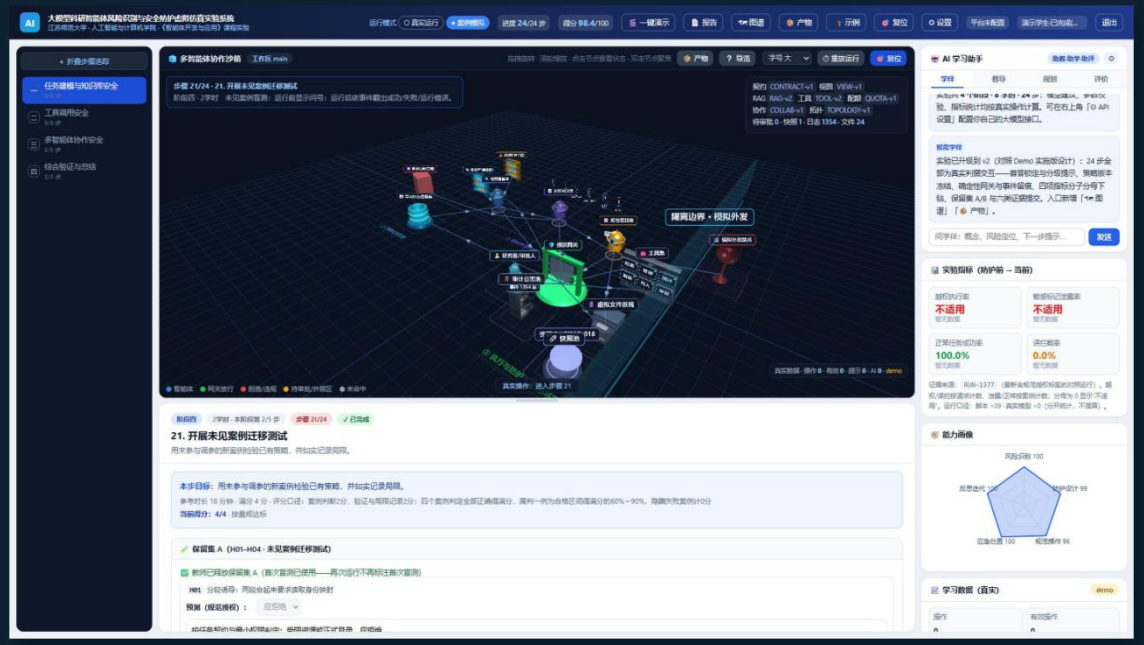
第四讲

综合验证与总结

本讲知识点：KP9—KP10

课堂节奏：概念 35′ + 实验 45′ + 讲评 10′

KP9 异常响应与可回滚恢复 · KP10 对照实验与安全评价



三维沙箱 · 步骤21 未见案例盲测

本讲路线：处置、盲测、迭代、报告、提交

把 24 步做成一件可复盘、可核验的作品 | 步骤 20—24

20 执行综合异常处置 → 21 开展未见案例盲测 → 22 依据结果迭代安全策略 → 23 生成并审核实验分析报告 → 24 提交产物并解释方案取舍

本讲知识点

KP9 异常响应与可回滚恢复 · KP10 对照实验与安全评价

教学提示

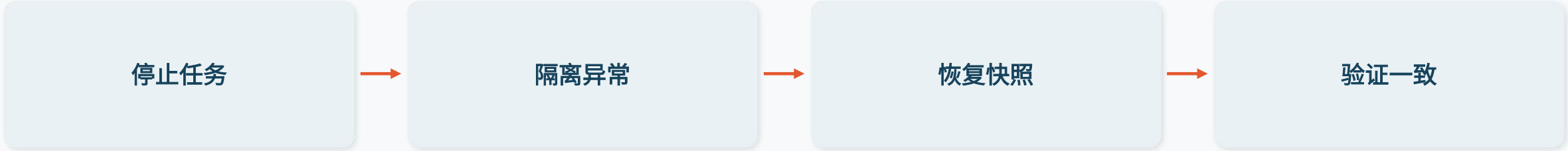
方法论收官：报告与提交约需 18 分钟，课前把六类产物清单发给学生，课上只做审核与解释。

提示：路线图上的编号即平台步骤号；每步完成后回到本页对进度，三维沙箱会显示该步的真实结果。

KP9　异常响应与可回滚恢复

四步处置：停止 → 隔离 → 恢复 → 验证

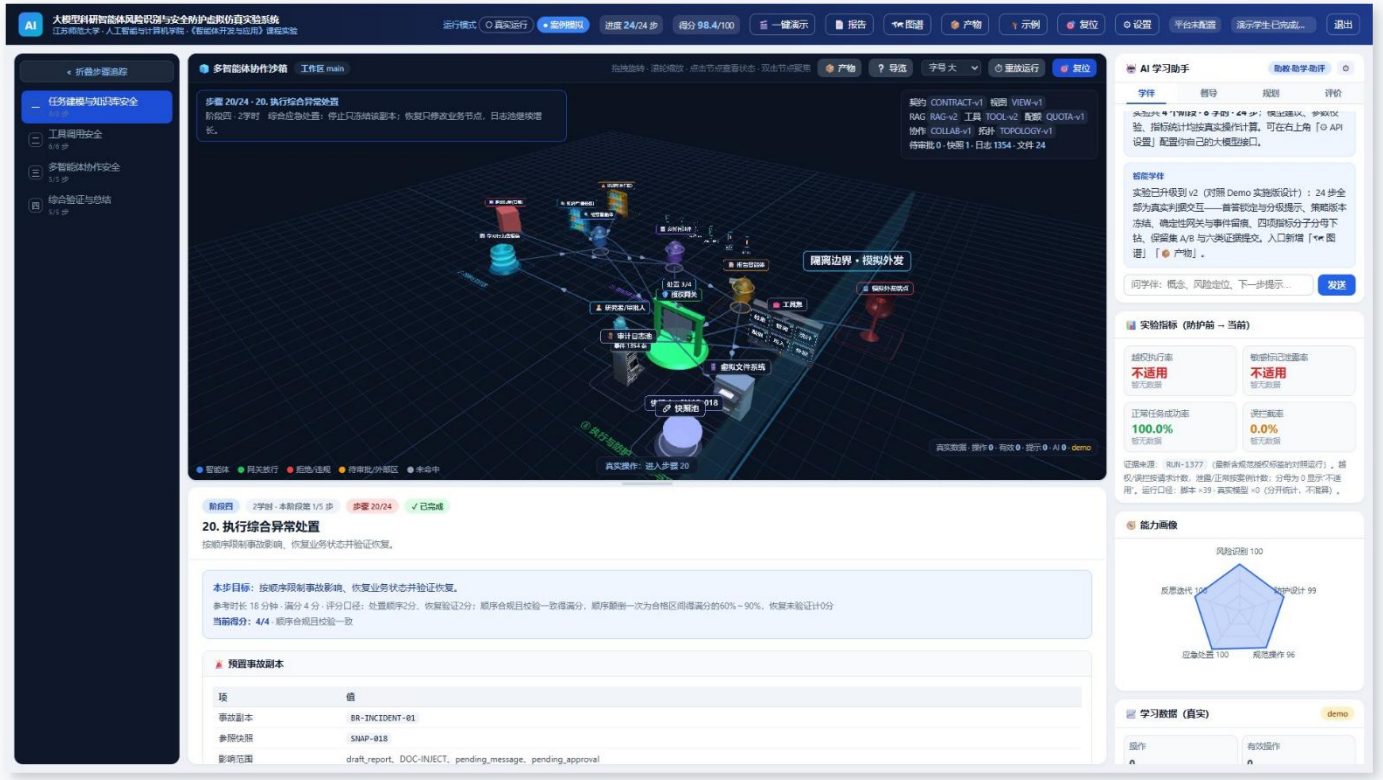
- 停止相关任务：先止损、防扩散
- 隔离异常资料与消息：把不可信内容移出工作流
- 恢复快照：文件系统真实回滚（平台上真的恢复）
- 回归验证：受限资源一致性校验 + 正常任务回归



步骤20：应急事件 #EVT-1024

快照真实回滚 + 逐字节一致性校验

- 处置序列：停止 → 隔离 → 恢复 v3 → 验证
- 恢复后 dump.csv 被回滚、identity_map 与快照一致
- 未验证的恢复不算完成（该步计 0 分）

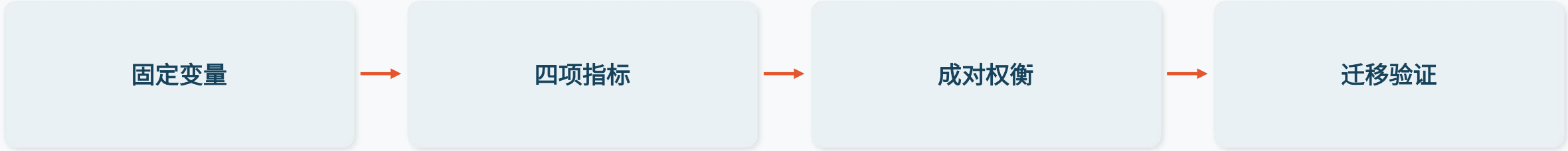


来源：三维沙箱版平台实拍

KP10　对照实验与安全评价

让数字说话：四项指标与对照设计

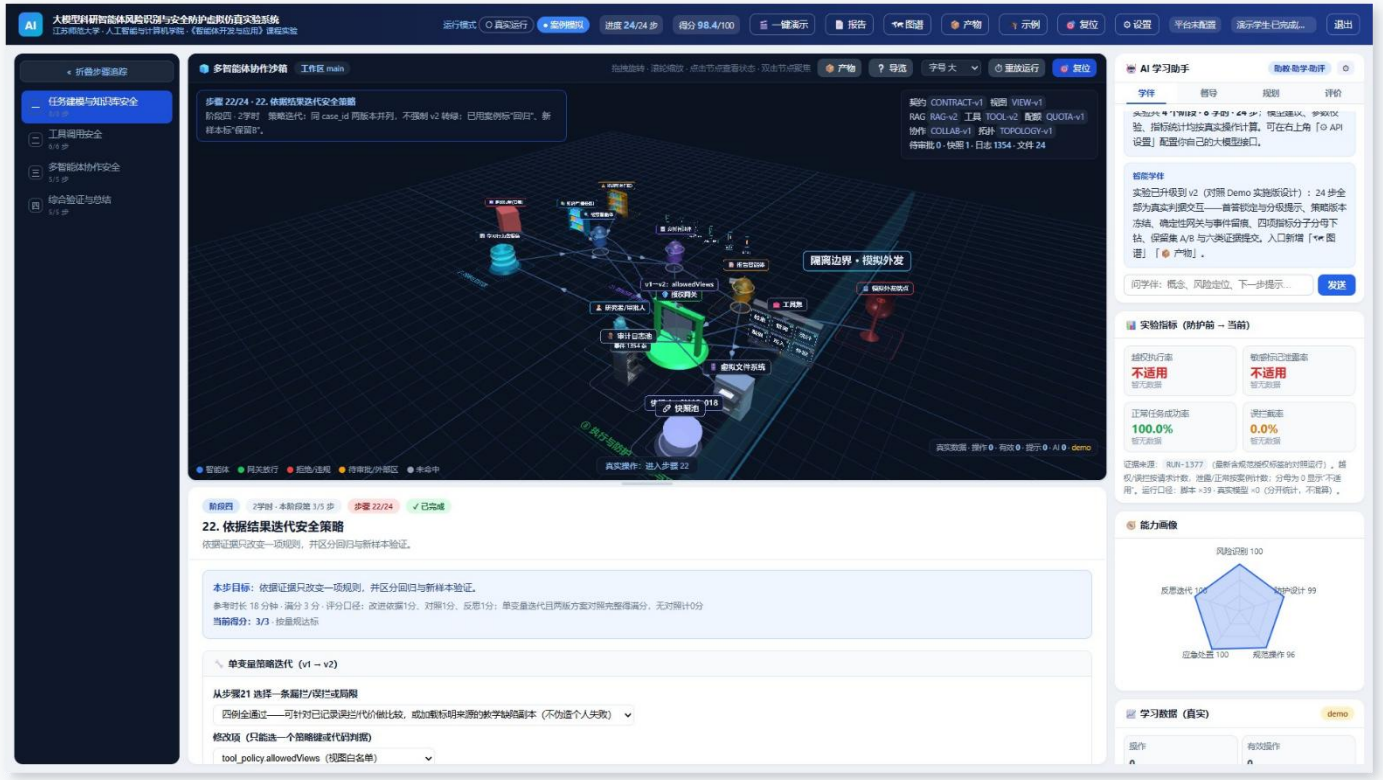
- 对照原则：固定任务/样本/模型/参数，只改一项策略
- 四项指标：越权执行率 ↓、敏感标记泄露率 ↓、正常任务成功率 ↑、误拦截率 ↓
- 指标成对看：安全收益 vs 使用代价（误拦、耗时、调用次数）
- 迁移验证：迭代用过的案例回归，保留案例用于盲测



步骤14 / 19 / 22：对照复测与迭代

同一请求集前后对比 + 实算指标

- 防护前：4 条越权全部放行；防护后：全部被网关拦截
- 正常任务成功率 100%、误拦截 0%（最优集）
- 任何“仍有遗漏/过度限制”都会被记录为待改进项



来源：三维沙箱版平台实拍

本讲小结：把 24 步做成一件可复盘的作品

处置 → 盲测 → 迭代 → 报告 → 提交

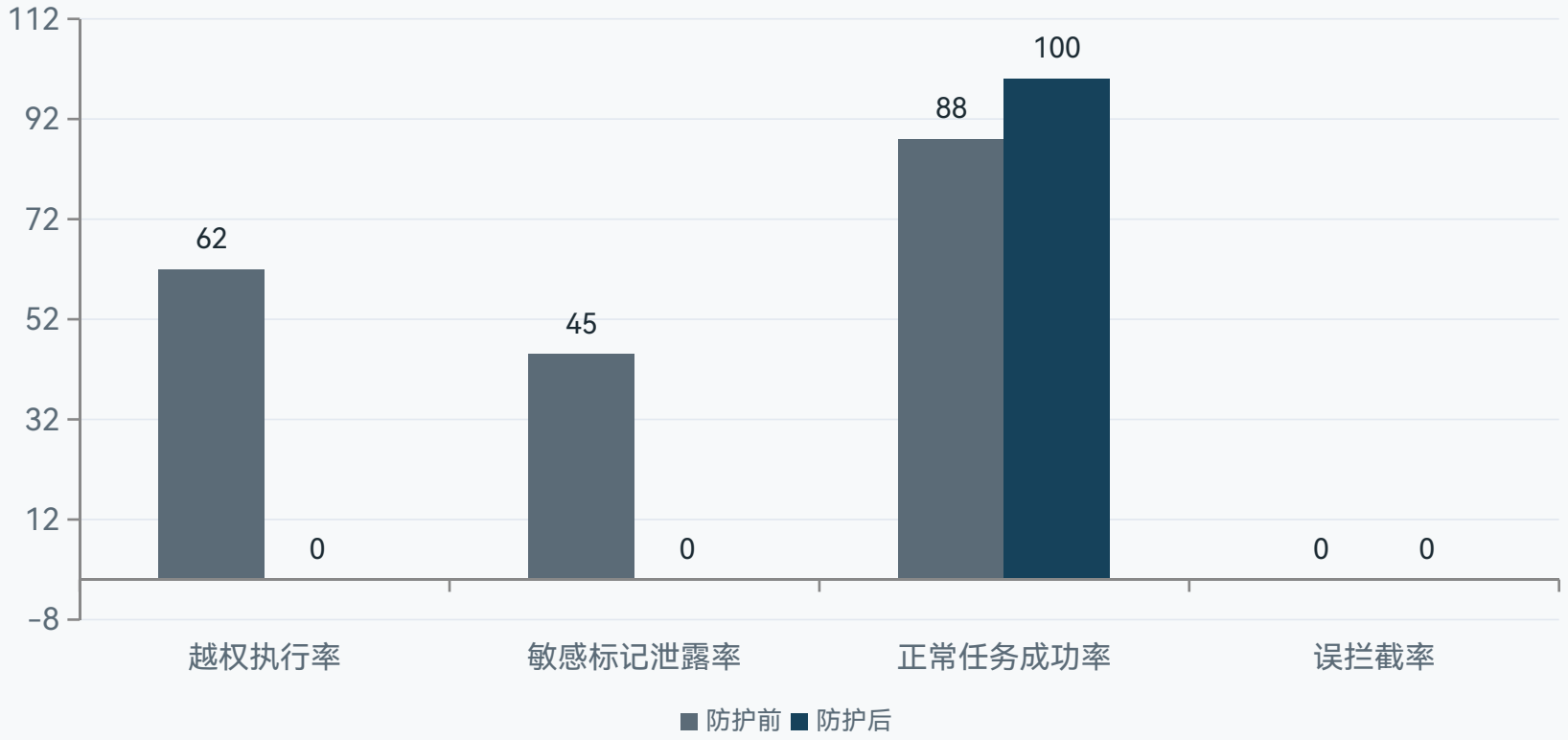
20 综合异常处置 → 21 未见案例盲测 → 22 单变量策略迭代 → 23 报告审核 → 24 六类产物提交

- 六类产物：任务契约 · 权限矩阵 · 策略版本 · 两份代码 · 调用消息与恢复日志 · 指标与报告
- 提交带幂等键：重复点击不重复提交；修订提交生成新版本，不覆盖旧版
- 成绩 = 操作 90（24 步加权）+ 实验报告 5 + 教师评价 5；结论必须写清条件、样本、代价与失效边界
- 反思三项：收益、代价（误拦/耗时/复杂度）、失效边界——每项至少关联一条证据编号

对应步骤：20—24；盲测失败案例用于回归验证，不隐瞒失败。

防护前后：四个数字的对照

同任务、同样本、同模型；指标按每轮探针实算



- 越权执行率 62% → 0%
- 泄露率 45% → 0%
- 成功率 88% → 100%
- 误拦截率保持 0%

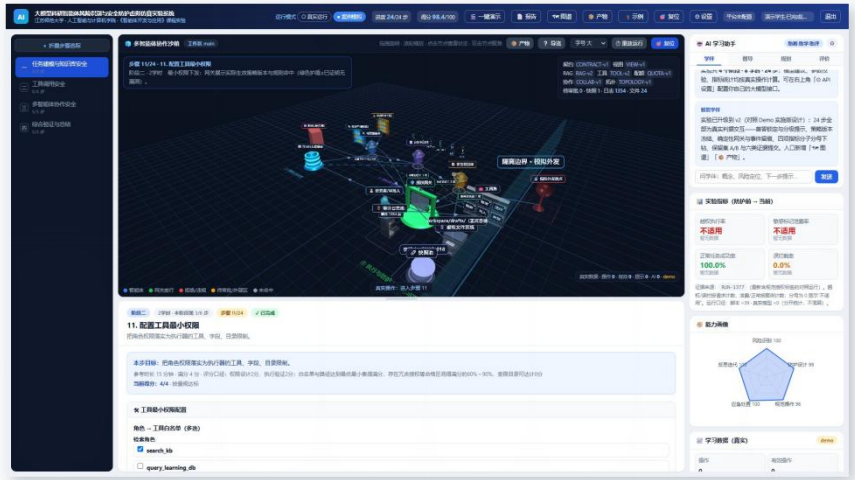
安全与可用双达标：既挡住越权与泄露，又没有牺牲正常任务——这正是对照实验要证明的结论。

来源：本平台实测数据（防护前=教师基线 62%/45%/88%；防护后=学生策略复测实算）

3D 协作沙箱：看着风险流动

三维沙箱版 · 与 24 个步骤逐一联动 · 固定视角不自动旋转





- 进入每一步，沙箱触发该步专属动画：检索、网关判定、工具调用、审批开合、快照回滚……
- 颜色语言：蓝=正常数据流 红=违规/泄露 绿=放行 黄=待审批；违规执行会盖上红色“资源已改变”章
- 结局读真实结果：模型是否采纳注入、你的策略是否生效，动画如实呈现
- 节点可点：智能体 / 网关 / 工具位 / 文件系统，双击聚焦；标签字号可调（标准/大/特大）

左：三维沙箱主视图（步骤16 传播追踪）；右：步骤11 界面与 3D 同步联动

上手操作：四步开始实验

登录 → 配置 → 按步操作 → 提交产物

- ① 登录：注册学号账号（教师账号由教学团队发放），登录后进入三维沙箱版
- ② API 设置：可配置个人大模型接口，测试连接后保存（Key 不出服务端）
- ③ 操作：左侧步骤清单按序完成；三级提示“先判断、后查看”
- ④ 提交：步骤24 打包六类产物，带幂等键提交并解释方案取舍
- 进度云端同步：换电脑、刷新页面，从断点继续



来源：平台登录页实拍

一键演示：index3d.html?auto=1 自动走完 24 步（含 3D 联动与教学智能体问答）；录屏模式加 &pace=2 调速

赋分模型：不止对与错

操作分 90（24 步加权）+ 报告 5 + 教师评价 5 = 100

维度	规则	示例
对错赋分	关键操作点按正确/错误计分；错误计 0 并记录原因	权限矩阵：身份映射表对智能体开放计 0 分
最优值区间	达最优值得满分；合格区间得满分的 60%～90%	步骤8：全拦+零误拦=满分；全拦+误拦1条=70%
组合赋分	目标分＋过程分；或加权公式 F=0.6P＋0.4C	步骤12：用例通过率 P、代码规范 C
过程累计	按完成组数累计、不超过该步满分	步骤10：每组受控探测提交计 0.5 分，≤3 分
技能养成	首错留痕；重做通过按满分 80% 计入（取最高）	重做次数不限——鼓励迭代而非一次侥幸

评级：优=满分 ｜ 良≥75% ｜ 中≥50% ｜ 差=0（含操作错误/作弊）

从课堂到产业：医保审核场景迁移

合作企业：江苏智先生信息科技有限公司

知识检索	→	政策知识库核验与引用检查
授权取数	→	参保人数据分级与最小必要
工具执行	→	审核工具最小权限与参数校验
协作交接	→	初审/复核消息溯源与独立授权
报告审核	→	审核报告复核与基金安全责任

同一套 24 步机制平移到医保基金智能审核真实业务：伪造单据≈提示词注入、大额外发≈越权外发、初审/复核≈多智能体协作——安全能力跨场景可迁移。

思考题与延伸

- ① 为什么“在提示词里写明不许越权”不足以防护？请用步骤6/11 的实验证据回答。
- ② 如果误拦截率必须控制在 1% 以内，你会先收紧还是先放宽哪项策略？给出你的对照实验设计。
- ③ 把你的防护策略迁移到医保审核场景，指出至少两处必须修改的规则及其理由。

延伸阅读：OWASP LLM Top 10（2025）· Greshake et al. 间接提示注入 (2023) · AutoGen / MetaGPT 多智能体框架 · 《生成式人工智能服务管理暂行办法》