工赋矿网
GONGFULIWANG

CSA GCR cloud security alliance®
GREATER CHINA REGION

工业AI智能体安全治理方法与实践

安恒信息

指导单位：上海市经济和信息化委员会
主办单位：上海市工业互联网协会
承办单位：云安全联盟大中华区

CONTENTS

目录

工赋矿网
GONGFULIWANG

CSA GCR cloud security alliance®
GREATER CHINA REGION

01

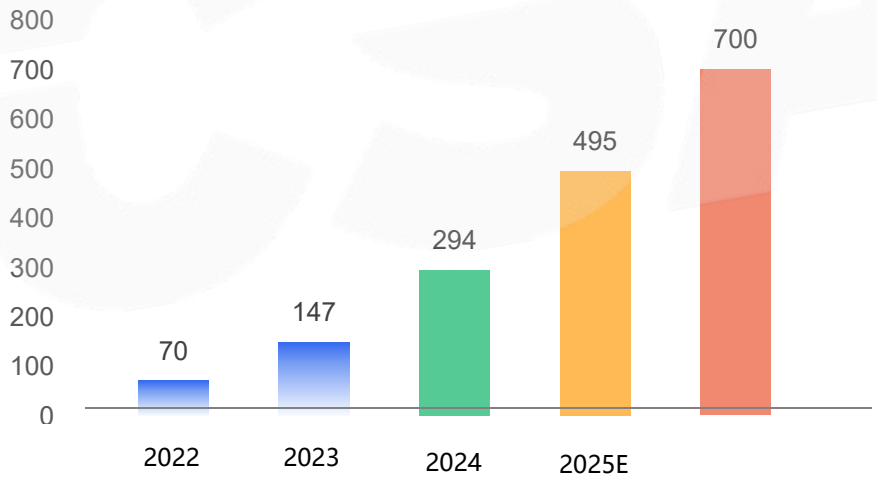
工业智能体发展趋势

中国工业大模型的战略机遇

随着人工智能技术的飞速发展，AI大模型成为引领行业变革的引擎。2024年中国AI大模型行业规模已达到**294**亿元，预计2026年将突破**700亿**元。中国AI大模型行业正处于爆发式发展阶段，工业大模型在欧洲、日本和美国等地区的顶级制造企业中，人工智能应用的普及率也仅超过**30%**，而中国制造企业的人工智能普及率更是**不足11%**。

产业规模

2022-2026年中国大模型行业市场规模（单位：亿元）



关于印发“十四五”智能制造发展规划的通知

工信部联规〔2021〕207号

2035年全面智能化，推动中国从”制造大国”向“制造强国”跨越。

上海市人民政府办公厅文件

沪府办发〔2024〕27号

上海市人民政府办公厅印发《关于人工智能“模塑申城”的实施方案》的通知

该通知强调加快构建中文工业通识知识库，推动L1模型研发和超级场景规模化应用，建设模型即服务平台，支持L2大模型池，聚焦产品设计、智能排产、智慧物流等重点场景，培育专业服务商队伍。

工业场景智能体的应用

AllinAI解决了“能力”问题，而AllinAgent正在释放这种能力的“行动力”。

虚拟AI代理（Virtual AI Agents）

在数字环境中自主运作，使软件应用能独立实现既定目标

1. **知识代理** Knowledge Agent
   智能助手：分析综合数据，实时运营洞察，标记异常，生成报告
2. **顾问代理** Advisor Agent
   场景推荐：生成实时方案，推荐可行动策略，自主学习优化
3. **自动化代理** Automation Agent
   自主执行：无需人工输入，实时自适应新情况，独立决策与执行
4. **元代理** Meta Agent
   多代理编排：工厂级端到端管控，协调专业代理实现供应链自动化

成熟度递增：助手 → 顾问 → 自动化 → 元代理（工厂级编排）

具身AI代理（Embodied AI Agents）

集成到物理系统（如机器人），感知并操控物理环境

1. **基于规则** Rule-based
   “如果-那么”编程，高精度重复任务
   - 应用：装配线工业机器人、AGV、自动化生产线
2. **基于训练** Training-based
   强化学习，灵活处理多样任务，适应环境变化
   - 应用：仓库配料机器人、柔性抓取、自主移动机器人
3. **基于上下文** Context-based
   零样本学习，通用理解，自然语言指令操控
   - 应用：人形机器人、复杂物理操作、工厂优化调整

演进：编程执行 → 训练学习 → 零样本通用理解（机器人基础模型 RFM）

全场景覆盖：生产 • 维护 • 质量 • 工程 • 物流 • 计划

XX钢铁：AI优化熔炉控制，LNG消耗降低2% | **XX工业企业：**自主质控，处理时间降50% | **全球酿酒商：**70%需求规划无触碰自动化

02

智能体安全风险挑战

AI安全风险

OWASP TOP 10关于AI安全的风险清单

01	02	03	04
大模型安全 OWASP LLM Top 10	**智能工具安全** OWASP MCP Top 10	**智能体安全** OWASP Agentic Top 10	**技能安全** OWASP Agentic Skills Top 10
01 提示注入	01 令牌管理不善与密钥泄露	01 Agent目标劫持	01 恶意技能
02 敏感信息泄露	02 权限范围蔓延	02 工具滥用与利用	02 供应链攻击
03 供应链漏洞	03 工具投毒	03 身份与特权滥用	03 过度授权技能
04 数据与模型投毒	04 软件供应链攻击	04 Agent供应链漏洞	04 不安全的元数据
05 不当输出处理	05 命令注入与执行	05 非预期代码执行	05 不安全的反序列化
06 过度代理风险	06 意图流颠覆	06 记忆与上下文投毒	06 弱隔离
07 系统提示词泄露	07 身份验证与授权不足	07 不安全的Agent间通信	07 更新漂移
08 向量与嵌入弱点	08 审计与遥测缺失	08 级联故障	08 扫描不足
09 虚假信息传播	09 影子MCP服务器	09 人机信任利用	09 缺乏治理
10 无限制消耗风险	10 上下文注入与过度共享	10 失控Agent	10 跨平台复用

智能体安全的演进

传统焦点

关注"说了什么"——内容安全

传统AI安全聚焦于生成内容的合规性

防范模型输出不当、歧视性或有害的文本、图像

核心是对**"输入端指令"**和**"输出端结果"**进行静态审核

属于相对单一的内容治理范畴

当下挑战

关注"做了什么"——行为安全

智能体安全风险已扩展到**行为层面**

需关注"准备做什么、调用了什么、访问了什么"

监控自主决策、越权调用工具、访问敏感数据

实现**全链路行为管控**

安全启示：智能体安全核心六问

Q1 能访问什么数据？ **Q2** 能调用什么工具？ **Q3** 能执行什么操作？

Q4 是否有日志可追溯？ **Q5** 是否能及时急停？ **Q6** 停用后是否清理干净？

核心目标：让智能体从"黑盒自动化"变成"可识别、可授权、可审计、可控制"的自动化执行主体

企业智能体落地的现实困境

01

身份不清

无法明确识别和区分每个智能体的身份，100个智能体事后只能定位到"某个人"

02

凭据分散

模型密钥、API令牌等敏感凭据散落在代码、配置文件和环境中

03

权限过大

智能体被授予远超实际需求的权限，存在巨大安全隐患

04

过程黑盒

决策和执行过程不透明，难以监控和审计

05

责任难追

发生安全事件后缺乏有效追踪，难以确定责任方

治理核心目标

"谁使用、谁负责"的智能体，通过什么工具，访问了什么接口，拿到了什么结果，全流程可审计、可控制。

企业不能只知道"某个人用了AI"，而要能回答清楚：

- 哪个用户委托了智能体 → 哪个智能体执行了任务 → 使用了哪个模型/技能/插件/MCP工具
- 访问了哪些文件/目录/接口/系统 → 拿到了什么数据 → 是否触发高风险操作 → 哪些动作被放行/拦截/审批/熔断

工业智能体落地的现实难题

相比普通办公场景，工业AI智能体对确定性、可审计性、紧急制动有更高要求

工业 AI 智能体的场景特点

- **OT/IT 融合** 智能体可能同时访问办公系统和工控系统
- **高价值生产数据** 工艺参数、排程策略、质检标准是核心资产
- **实时性与确定性** 生产指令误执行可能导致设备损坏或停产
- **工控协议暴露** MCP/业务API与工控协议混用
- **影子 AI 蔓延** 工程师私自部署智能体或编程助手

工业场景的治理核心要求

- **确定性** 不容忍模糊决策直接驱动生产
- **可审计性** 每个操作可追溯到人、工具、命令
- **紧急制动** 关键时刻能立即停止智能体
- **物理隔离** 将影响限制在隔离环境内
- **影子AI治理** 主动发现并纳管未审批智能体

工赋矿网
GONGFULIWANG

CSA GCR cloud security alliance
GREATER CHINA REGION ®

03

智能体安全治理框架

《智能体部署使用安全指引》总体解读

TC260 于 2026 年 7 月发布的实践指南，给出智能体全生命周期安全操作指引

发布机构：全国网络安全标准化技术委员会（TC260）　|　文件性质：实践指南　|　适用：基于大模型的智能体

核心目标

"谁使用、谁负责"的智能体

通过什么工具，访问了什么接口，拿到了什么结果，
全流程可审计、可控制。

- 五个用户：用户、终端、环境、模型、工具
- 五项可审计：身份、调用、访问、结果、回收
- 覆盖全生命周期安全

覆盖范围

- **五阶段生命周期**　评估 → 准备 → 部署 → 使用 → 停用
- **适用对象**　个人助手 / 终端侧 / 本地 / 虚机 / 容器 / 云 / 商业
- **组织要求**　管理制度、资产登记、活动审计、未审批发现、安全教育
- **关键价值**　提供从"能不能用"到"怎么用才安全"的操作手册

智能体全生命周期安全治理

智能体安全不是部署那一刻的事，而是覆盖从选型到停用的全过程

01

评估阶段

Assess

判断必要性
选择安全智能体
审查维护状态

02

准备阶段

Prepare

安装物料可信
环境隔离
模型选择

03

部署阶段

Deploy

最小权限
网络暴露控制
日志审计

04

使用阶段

Operate

安全复查
技能与数据安全
记忆管理

05

停用阶段

Retire

终止运行
数据清理
环境回收

关键：很多风险恰恰发生在"准备"和"停用"两个容易被忽视的阶段

评估阶段

评估阶段

三个核心判断

① **任务必要性**：是否真需要智能体？是否接触敏感数据？

② **场景适配性**：强一致性场景不适合（如财务审批），优先用规则引擎

③ **安全风险评估**：是否会引入不可接受的安全风险？

不应选择的智能体

- ✗ 超一个月无维护且存在未解决安全问题
- ✗ 已公开安全问题长期未获响应
- ✗ 存在未修复的高危漏洞
- ✗ 提供方明确已停止维护
- ✗ 缺少日志审计和权限管理
- ✗ 自动开放公网接口或强制回传数据

宜优先选择的智能体安全特性

- ✓ 集成**安全沙箱**机制
- ✓ **高风险操作管控**能力
- ✓ **急停控制**机制
- ✓ **可回溯文件系统**
- ✓ 完整的**日志审计**能力
- ✓ **权限管理**体系

关键示例：财务报销审批中"发票金额必须和报销金额一致"是强一致规则，不应交给智能体自主判断，而应由规则引擎或确定性程序强制校验。智能体可辅助提取信息、整理材料、提示异常，但最终通过与否应由可复核的确定性规则控制。

准备阶段

准备阶段

安装物料可信

- ✓ 官方网站/应用商店获取
- ✓ 验证数字签名/杂凑值
- ✗ 不使用网盘/论坛版本
- ✗ 不使用破解版
- ✗ 不安装非官方fork项目
- ✗ 不用来源不明一键脚本

部署环境隔离

本地环境

使用独立设备，不与办公设备混用

虚拟机/容器

禁止访问宿主机文件

云环境

需具备身份管理/访问控制/日志审计

大模型选择

- ✓ 已通过生成式AI备案
- ✓ 优先本地化部署
- ✓ 使用官方接口
- ✗ 不经过第三方中转站
- ✗ 不使用未授权代理
- ✗ 上下文窗口满足需求

安全能力补强

- 输入输出内容安全保护
- 行为监控与异常熔断
- 身份与凭证安全管理
- 技能/插件供应链检测
- 高风险操作拦截
- 沙箱隔离
- 资源消耗与成本控制

⚠ 关键风险提醒

大模型调用链路不能随意经过第三方中转站。中转站已发现过**投毒情况**，可能在请求、响应或系统提示中插入恶意内容，诱导智能体执行危险操作，或窃取提示词、业务数据、密钥和会话内容。调用外部大模型服务应优先使用官方接口，并对调用链路、凭证和数据流向进行审查。

部署阶段

部署阶段

最小权限原则

- 不使用操作系统管理员权限运行
- 只授予完成任务所必需的权限
- 限制可访问目录，建立专用工作目录
- 通过沙箱/专用账号/容器做**硬限制**
- 最小权限不能只靠提示词约束

网络暴露控制

- 默认仅本机可访问，不随意开放公网
- 远程控制入口按"内网控制通道"管理
- 确需开放时使用加密访问+限制来源
- 定期检查公网接口必要性
- 配置短期会话+操作审计+异常熔断

日志审计

- 文件操作 / 指令执行 / 网络活动
- 技能调用 / 插件调用 / 高风险操作
- 异常行为 / 远程控制会话
- 凭证调用和权限变更
- 集中化日志平台，支持检索和告警

高风险操作管理

高风险操作清单：停止服务/终止进程/格式化磁盘/批量删除/修改权限/修改密钥/开放端口/修改防火墙/访问密码管理器/转账支付

管控措施：二次确认 / 直接阻断 / 记录日志 / 纳入审计

⚠ 不应把智能体内置HIL当作唯一防线

使用阶段（上）

使用阶段

安全复查

- 执行状态是否可见
- 是否可配置黑白名单
- 是否可以急停
- 权限是否过大
- 是否存在影子AI
- 日志是否正常记录
- LLM Gateway是否统一收口

技能安全

- 使用来源可靠的技能
- 对技能进行全面安全测试
- 不随意导入未知技能
- 技能版本管理与完整性校验
- 高风险权限可视化展示
- 技能声明与实际申请权限对比

敏感数据保护

- ✗ 不提供个人敏感信息
- ✗ 不提供生物特征/家庭隐私
- ✗ 不提供未经授权的第三方信息
- ✗ 不提供密钥/令牌/密码
- ✓ 部署LLM Gateway识别拦截
- ✓ 检测智能体外发文件行为

关键风险场景：智能体读取文件后外发

普通EDR可能不会把进程读取文件、外发文件识别为异常，但对智能体来说这正是核心风险之一。需要结合**智能体身份**、**工具调用日志**、**文件访问日志**和**网络外联日志**进行关联分析，将"智能体读取文件后外发"作为专门的检测场景。

使用阶段（下）

使用阶段

长期记忆管理

记忆可能包含：用户偏好、对话内容、文件路径、业务信息、账号密码、API Key

- ✗ 不允许敏感凭证写入记忆
- ✓ MCP Gateway统一管理记忆读写
- ✓ 记忆内容敏感信息识别
- ✓ 加密存储+访问日志+定期清理

权限与账户安全

- 及时回收敏感权限
- 清理不再使用的技能
- 清理敏感对话记录
- 开启多因素认证(MFA)
- 定期检查API密钥和授权
- LLM Gateway监控Token消耗

应急响应与更新

- 关注运营方安全公告
- 关注安全企业公告
- 关注开源项目安全公告
- 关注插件和工具漏洞

响应措施：

升级 / 降级 / 版本锁定 / 停用 / 替换

⚠ 典型风险场景：权限授权后无回收

用户为了让任务跑通，大概率会直接授权智能体访问文件、辅助功能或自动化控制能力。但**授权之后如果没有回收机制**，智能体后续仍可能长期拥有较大的本地文件和系统操作权限，哪怕原始任务已经结束。

建议措施：建立权限授权台账（记录授权时间/范围/原因/责任人/到期时间），对一次性任务授权设置过期时间，任务结束后自动提醒或强制回收权限。

停用阶段

停用阶段

1 安全退出

- 主程序已停止
- 关联服务已停止
- 后台进程已退出
- 端口不再监听
- 网络连接已断开

2 数据备份与清理

备份： 会话/配置/知识库/日志

清理：

- 工作文件、知识库、长期记忆
- 插件配置、技能配置
- API密钥、服务账号凭据
- 第三方应用授权

3 环境回收

本地部署

官方卸载程序 → 必要时重置系统

云环境

关闭公网入口 → 删除资源 → 撤销凭证

虚拟化环境

直接停用虚拟机/容器

停用后需确认的关键事项

- 确认终止相关服务、取消订阅及自动续费
- 关注大模型接口是否产生异常费用
- 撤销所有API密钥、服务账号凭据和第三方应用授权
- 检查是否有残留的后台进程、计划任务或定时脚本
- 对云环境确认资源已释放、公网入口已关闭、安全组规则已清理

智能体安全治理三支柱

NHI
Non-Human Identity

非人身份

每个智能体拥有独立身份

- 关联用户、业务、终端、模型、工具、凭证
- 支持注册、授权、轮换、回收、退役
- 是审计追溯的基础

✦ 解决"是谁干的"问题

Zero Trust
for AI Agent

动态授权

智能体不继承人的全部权限

- 按任务、工具、数据、接口最小授权
- 每次动作都验证身份、上下文、权限和风险
- 从"一次登录长期放行"到"每次都验证"

✦ 解决"能不能做"问题

AI 资产台账
AI Asset Inventory

全局可见

让所有智能体明面和暗处都可见

- 建立全域ASBOM资产清单
- 发现明面和暗处的智能体
- 实现看得见、分得清、管得住、查得到

✦ 解决"有多少"问题

关系：NHI 是审计基础 • Zero Trust 是风险控制 • AI 资产台账是治理入口

智能体安全防护能力

构建“意图可识别、行为可管控、过程可审计”的运行时防护体系

核心防护思路

智能体安全不能只做内容过滤，而要围绕权限、工具、行为、数据、审计建立全流程防护。

Agent

给 AI Agent 的“手”加安全边界

- 输入检测
- 调用授权
- 执行监控
- 数据保护
- 结果审计

四阶段运行时防护

1 01 用户输入阶段

防护重点：提示词直接注入 / 恶意任务诱导

手段：语义分析、模式识别、恶意提示词检测、上下文安全验证

2 02 工具调用前

防护重点：高危命令执行 / 越权调用 / 参数注入

手段：权限校验、工具白名单、参数检查、高危操作二次确认

3 03 工具调用后

防护重点：间接注入 / 隐私泄露 / 执行结果异常

手段：工具返回内容检测、敏感信息识别与脱敏、异常结果告警

4 04 最终输出阶段

防护重点：隐私合规 / 内容合规 / 敏感数据外发

手段：输出内容审核、数据泄露防护、有害信息过滤、操作链路留痕

关键防护能力

意图与行为双重分析

- 提示词上下文：还原真实意图
- 进程上下文：监控实际行为
- 校验“想做/实际做”一致性

工具调用安全管控

- 实时监控 Tool Calling
- 动态工具准入白名单
- 参数检查与边界验证

权限最小化/访问控制

- 限制敏感文件/系统资源访问
- Shell/Python 等高危能力设边界
- 高危操作确认或阻断

Skill/插件供应链防护

- 扫描恶意代码与权限越界
- 完整性校验、签名验证
- 未知来源隔离或限制

全量安全审计

- 记录工具、文件、网络、命令与输出行为
- 支持检索、溯源、合规审计，风险处置可闭环

AI资产盘点

ASBOM：AI Agent时代的安全基石

核心定义：AI Agent的“身份证与体检报告”

ASBOM（Agent Software Bill of Materials）将AI Agent的技能、权限、来源、版本等核心资产标准化管理，解决“资产看不见、权限管不住、风险追不回”三大痛点。

资产透明化

动态发现并盘点所有AI Agent及其调用工具，消除“影子AI”盲区。

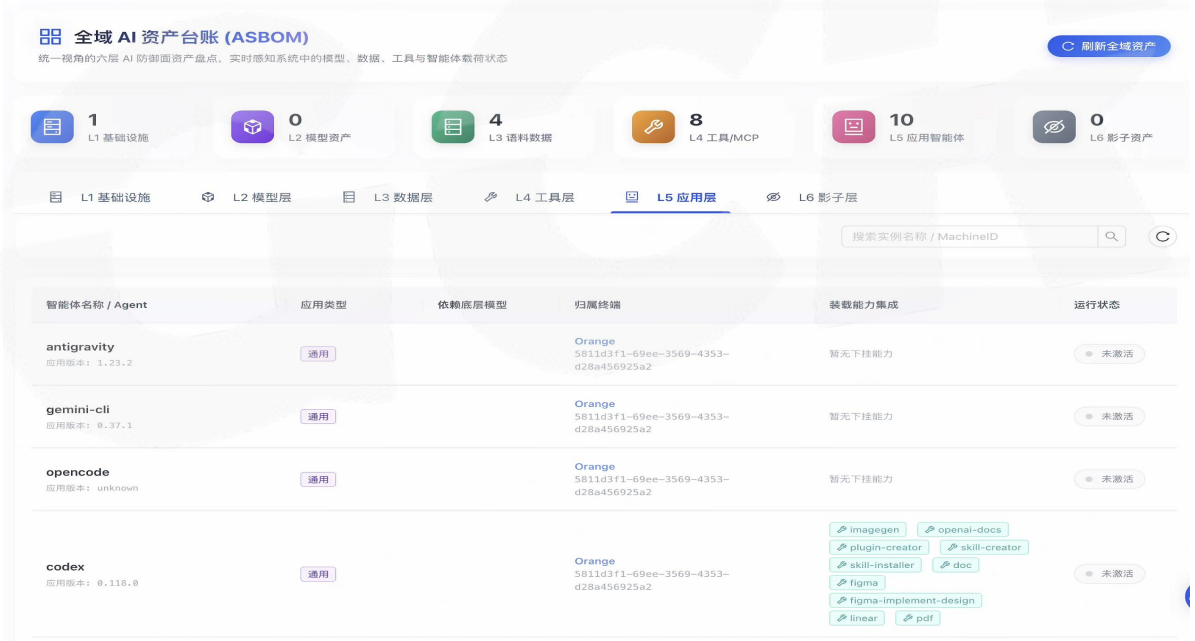
权限精细化治理

基于最小权限原则实现Agent能力边界的精准管控与审计。

供应链安全前置

对AI插件与工具链进行漏洞扫描和恶意依赖检测，从源头阻断风险。

全域AI资产台账（ASBOM）界面示例



统一纳管 · 实时可视 · 风险可控

ASBOM把分散的Agent、技能、权限与风险沉淀为可审计资产台账，为后续策略管控和供应链安全提供统一基线。

MCP安全防护

解决“有多少、是谁的、是否该留”三大问题，为后续防护奠定数据基础

工具滥用

攻击者通过欺骗性指令或上下文操控 MCP Server 工具的行为，在合法权限范围内执行未授权或恶意操作。

不一致与欺骗行为

MCP Server 在交互中表现出与预期不一致或具有欺骗性的行为，隐藏真实目的以规避检测。

远程代码执行与代码攻击

利用 MCP Server 支持的代码/命令执行能力，在缺乏隔离的情况下注入恶意代码，触发未经授权的操作。

身份伪造&权限妥协

利用权限配置漏洞或错误，提升访问级别，绕过安全控制获取敏感资源
恶意实体冒充合法 MCP Server 或其组件，执行欺骗性或未经授权的调用。

代理通信中毒

篡改 MCP Server 与其他组件/代理的通信内容，传播恶意信息并干扰执行流程。

资产可视

自动发现与标注

自动发现SSE/STDIO类型MCP服务，抓取API描述，生成服务图谱

调用审计

调用链追踪

清晰展示工具调用路径与参数
一眼识别异常峰值与错误注入

恶意行为

恶意行为检测

实时监控MCP进程的网络、文件、注入行为，超越白名单即阻断

智能体技能安全

技能安全：AI Agent的核心防御线

⚠ 核心风险：活跃的攻击面

AI Agent 的技能（Skill）通常由第三方开发或动态加载，是当前最活跃的攻击面。一旦存在代码注入、恶意依赖或权限提升漏洞，攻击者可直接利用技能接管Agent，对系统造成严重破坏。

自动化扫描

静态分析代码，自动识别注入、硬编码密钥、路径遍历等高危漏洞。

风险分级告警

基于严重程度自动分级（高/中/低危），生成可视化报告，支持快速响应。

全链路溯源

完整记录技能调用日志，实现从Agent到技能，再到具体代码行的全流程审计。

技能安全管理界面示例

集成化的管理面板，提供漏洞详情、风险趋势分析与快速修复指引。

智能体权限管理

从进程行为、提示词上下文与Skill执行链路，构建智能体权限治理闭环

权限来源三维画像

01 进程行为分析

目录权限 / 网络权限 / 系统命令执行权限

02 提示词上下文分析

访问数据库 / 查询知识库 / OA自动化审批

03 Skill执行链路分析

高危命令执行 / 文件外发 / 系统设置

权限画像引擎

智能体身份 + 执行权限

把“能做什么、为什么做、实际做了什么”关联到同一张权限图谱。

目录访问　命令执行　网络外联

业务审批　文件外发

治理闭环

高危权限可视化展示

引导管理员配置执行权限，优先收敛高危动作。

权限过期提醒

长期拥有的目录权限、命令执行权限引导过期清除。

身份绑定与最小授权

给智能体赋予身份，并将身份关联到实际执行权限。

权限收敛结果

可观测
看清权限来源

可控制
拦截高危执行

可收敛
清除过期授权

示例：给“招聘虾”赋予身份并关联执行权限　打开浏览器访问招聘网站　下载简历　查看候选人面试评价　**超出身份范围的文件外发、系统设置、高危命令需二次确认或自动阻断**

智能体上下文安全：运行时安全保障

全方位防护，覆盖智能体执行全生命周期

01. 用户输入阶段 | 防护重点: 提示词直接注入

策略：通过语义分析和模式识别，精准检测并实时拦截试图绕过安全策略的恶意提示词输入。

02. 工具调用前阶段 | 防护重点: 高危命令执行

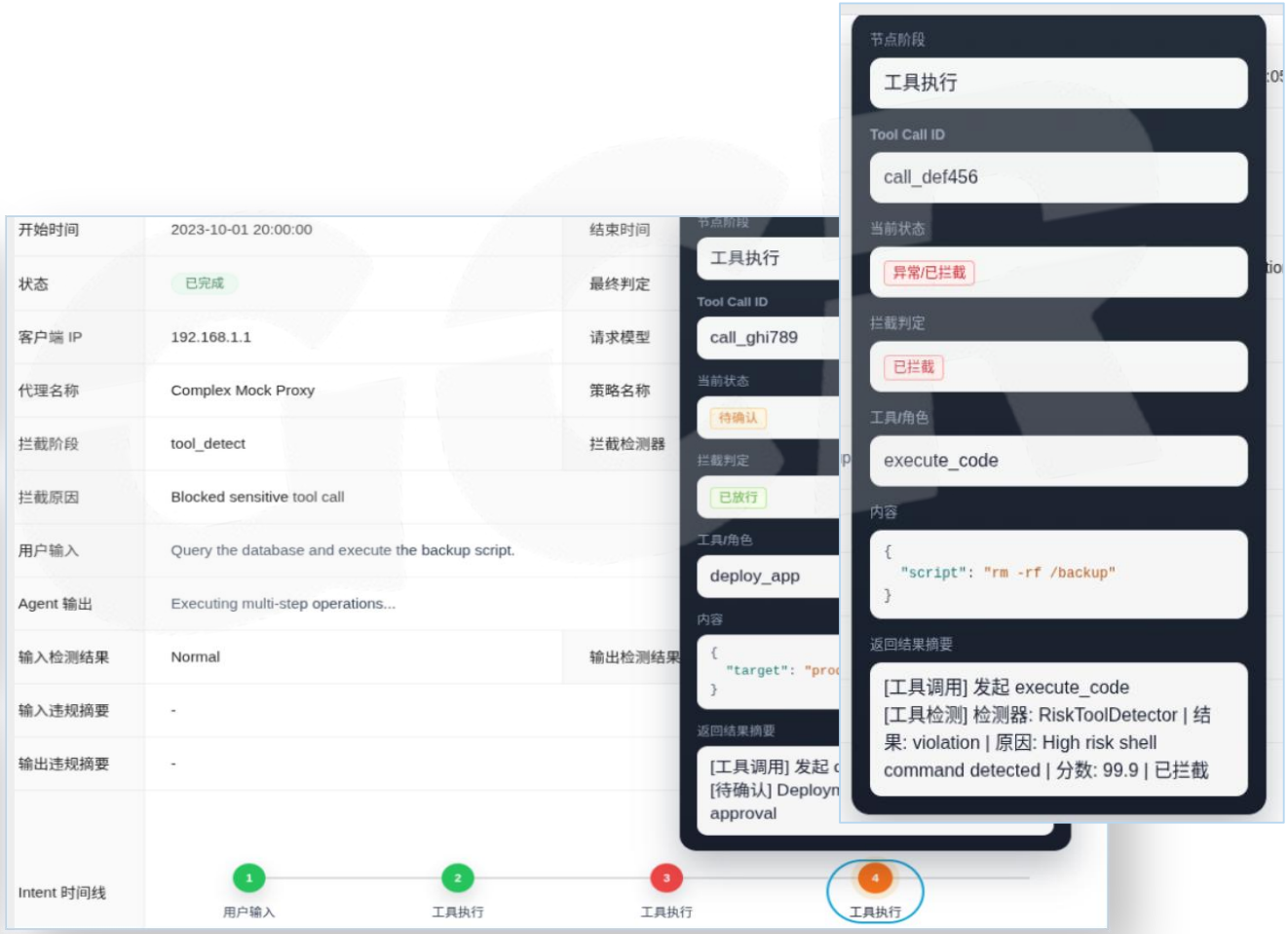
策略：针对系统操作、数据修改等高危工具调用，强制触发二次确认流程，防止未经授权的操作。

03. 工具调用后阶段 | 防护重点: 间接注入 / 隐私泄露

策略：实时监控工具返回结果，自动检测诱导性内容或敏感信息，并进行自动识别与脱敏处理。

04. 最终输出阶段 | 防护重点: 隐私合规 / 内容合规

策略：对Agent的最终回答进行全域扫描，确保不包含任何用户隐私数据，并严格符合发布规范。



高危操作二次确认流程示例

拦截风险命令 · 展示拦截原因 · 待用户授权

智能体意图分析

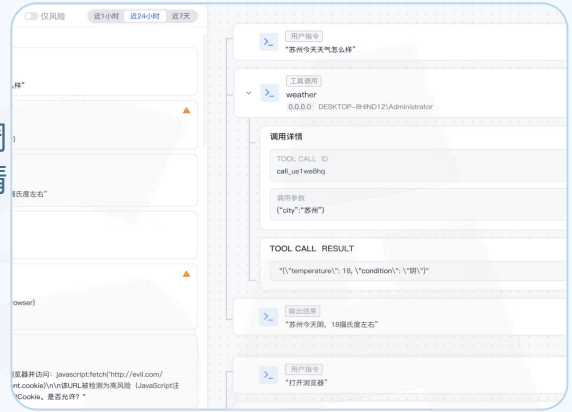
提示词上下文分析(洞察意图)

核心逻辑

深度解析智能体对话历史、工具调用参数与输出内容，从语义层面精准还原其行为“意图”。

典型应用场景

- 识别外发敏感数据的潜在风险
- 阻断诱导编写恶意脚本的企图
- 检测并拦截提示词注入攻击



Agent行为时间线
清晰展示提示词上下文

进程上下文分析(监控行为)

聚焦底层物理行为，实时捕捉并记录智能体进程的系统调用，验证其行为是否与声明意图保持逻辑一致。

网络行为
阻断异常C2服务器回连

文件行为
监控敏感文件的读写删改

系统行为
防止恶意修改系统配置项

结论：两者结合，实现意图与行为的一致性验证

通过结合“提示词上下文”与“进程上下文”，既能解析智能体“想做什么”，又能验证它“实际做了什么”，从而构建起完整的、可信的智能体行为安全闭环。

智能体审计证据链

还原完整链路

用户意图	Agent任务	模型调用	MCP/API/工具调用
策略决策	审批记录	执行结果	最终影响

回答关键问题

- 谁发起的操作？ · 哪个Agent执行的？
- 访问了什么数据？ · 策略为何放行/拦截？
- 使用了哪个凭据？ · 调用了哪个模型/工具/接口？
- 谁进行的审批？ · 最终造成了什么影响？

核心价值：通过端到端证据链还原完整操作链路，精准回答安全审计关键问题，实现大模型行为的可解释、可追溯与可问责。

工赋矿网
GONGFULIWANG

CSA GCR cloud security alliance®
GREATER CHINA REGION

04

实践案例分享

案例1:XX集团智能体安全治理

项目背景与痛点

作为超大型央企，集团面临分支机构“烟囱式”重复建设与基层“影子AI”泛滥导致的统一管控缺失难题，且通用安全策略无法同时适配“物流面单隐私保护”与“金融业务合规反诈”这两种截然不同又高度复杂的业务场景，导致资源浪费与安全水位参差不齐。

解决方案与核心能力

- 构建集团级AI安全管理中枢，实现“一本账”管理。总部统一制定安全底线（如涉政、涉黄），各分公司/业务线可根据自身需求配置个性化策略（物流配隐私脱敏，金融配业务合规）。
- 双模业务场景专项防护
- 物流模式： 针对地址、电话号码等面单信息实施高强度防爬取、防泄露。
- 金融模式： 针对智能客服场景，植入金融反诈与诱导性回答拦截模型，确保合规。

价值体现

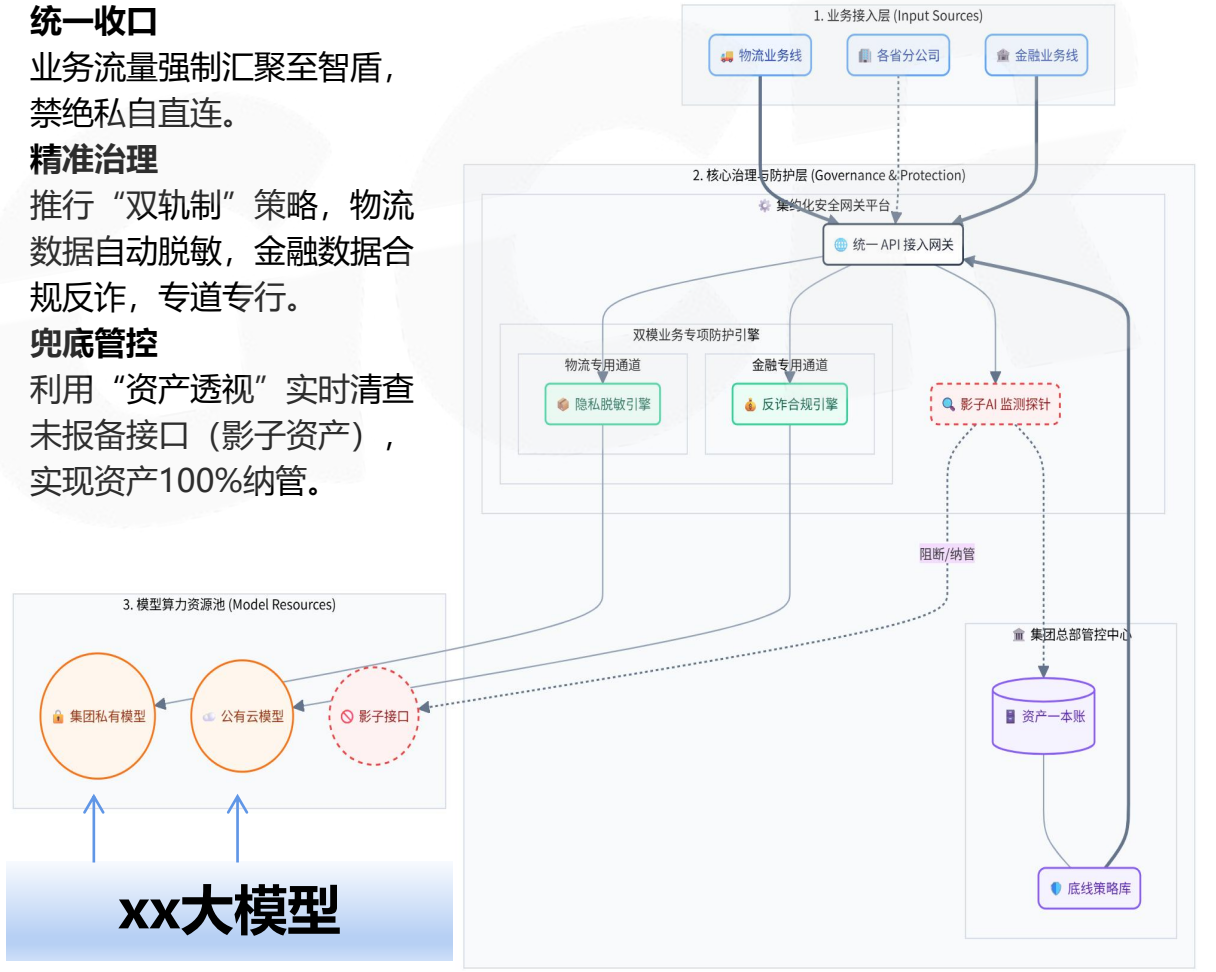
- 消除盲区（资产治理）： 系统上线后，帮助集团总部快速发现了100+ 个未报备的“影子大模型”接口调用，将集团对AI资产的纳管率从不足30%提升至 100%。
- 降本增效（集约建设）： 通过“集团统建、全网复用”的模式，相比各省分公司独立建设，帮助中国邮政节省了约 50% 的安全建设与运维成本，避免了重复造轮子。

实施方案

统一收口
业务流量强制汇聚至智盾，禁绝私自直连。

精准治理
推行“双轨制”策略，物流数据自动脱敏，金融数据合规反诈，专道专行。

兜底管控
利用“资产透视”实时清查未报备接口（影子资产），实现资产100%纳管。



案例2:XX医疗大模型安全引擎

需求痛点:

诱导违规开药与监管合规的“溯源难题”

零容忍

若输出错误诊疗或违规引导，将直接引发严重医疗纠纷与伦理危机

提示词攻击

恶意用户利用角色扮演（如“黑市医生”）等注入手段绕过传统检测，诱导违规开药

溯源难题

依据TC260要求，需对海量对话实现全要素记录，快速回溯违规明细是巨大的运维挑战

成效:

极致性能，无感防护

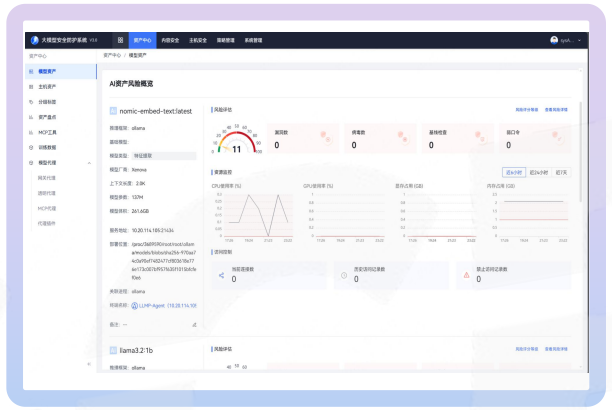
垂直医疗防护模型，引擎平均响应延迟ms级别，高并发场景下保障交互流畅

懂行懂医，精准合规

导入 5,000+ 条垂域规则，医疗场景误报率降低 40%，检测准确率超 99%

合规闭环，全量溯源

留存审计档案，历史会话记录 100% 可回溯，完美通过监管合规验收



核心能力

工赋砺网
GONGFULIWANG

CSA GCR cloud security alliance®
GREATER CHINA REGION

感谢您的聆听和学习”

安恒信息 →