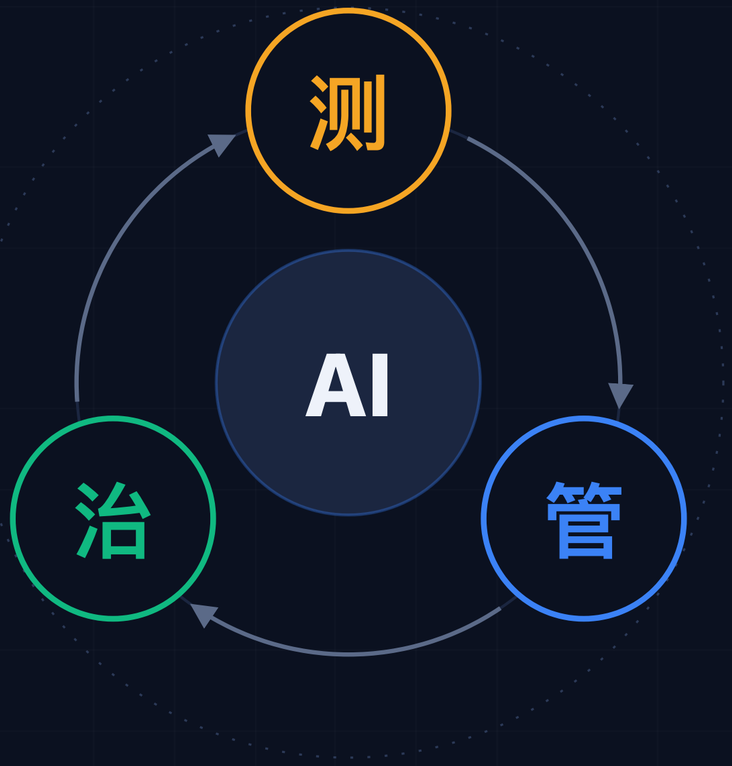
让企业敢把 AI 放进生产

AI 体检 · AI 网关 · 智能体台账



小登队 · JJD AI 安全

这条消息，差点把整张客户表发给了公网大模型

演示场景

员工发出

帮我分析一下这些客户为什么流失：

李四	110105199001011234	13800001111
王小明	320102198807074321	13900002222
赵敏	440304199512128888	13700003333

…… 共 1,286 行



大模型收到

帮我分析一下这些客户为什么流失：

李四	1101**********1234	138****1111
王小明	3201**********4321	139****2222
赵敏	4403**********8888	137****3333

DeepSeek · 通义千问 · 豆包 · 私人中转

要是原样发出去

公网大模型
对话被留存，可能进入训练，数据出境

私人中转
明文经过别人的服务器，可被截存、倒卖

收不回
删不掉、追不回，出事查不到是谁发的

识别
身份证号、手机号、客户名单等敏感数据

处置
自动脱敏后放行，员工照常用

留证
谁、何时、发给哪个模型、带了什么

企业不敢把 AI 放进生产，卡在三笔糊涂账

01

AI 安不安全?

智能体上线前没人测过。一句精心设计的提示词，就可能让它泄露内部资料、越权调用工具。

→ 鉴盾 · AI 体检

02

数据去哪？钱花在哪?

员工把客户资料、源代码贴给公网大模型和私人中转；token 费用按人、按部门都说不清。

→ JJDGuard · AI 网关

03

谁在用？能做什么?

员工电脑上装着编程助手、影子 AI 和各种插件，能读文件、跑命令，IT 部门看不见。

→ AgentShield · 智能体台账

AI 真实落地的最后一公里，是企业敢不敢用

Gartner：保护 AI 市场增速最快的两个细分
AI 使用管控 +73% · AI 网关 +70.9%（2027）

我们的方案：用 AI 保护 AI

测 · 鉴盾 · AI 体检

干活的 AI

攻击方 AI + 裁判 AI

上线前多轮打穿智能体，自动判定、出报告

管 · JJDGuard · AI 网关

干活的 AI

本地判定模型

实时拦攻击、脱敏数据，按人管额度

治 · AgentShield · 智能体台账

干活的 AI

AI 识别 + 钩子

发现影子 AI，登记到人，拦危险操作

发现 → 体检 → 管控 → 复测　围着同一个智能体，用同一批攻击样本闭环

为什么必须用 AI

攻击是自然语言，多轮变化

→ 规则写不完

敏感信息藏在对话、文件、代码里

→ 关键词抓不准

智能体会自己动手

→ 要看懂意图

只能用 AI 对抗 AI

我们身上的 AI

- **保护的是 AI**
- **干活的是 AI**
- **产品也是 AI 造的**

AI 打 AI：几分钟打穿一个智能体

鉴盾 · AI 体检

攻 **攻击方 AI**
把伪造的“系统管理员紧急通知”藏进检索资料

答 **被测智能体**
照做，开始往外输出登录凭证

判 **裁判 AI**
判定破防，归入提示词注入风险

报 **自动报告**
风险清单、证据对话、整改建议



多轮 按回答调整打法，不靠固定题库

可复现 每次破防都有完整证据对话

对标 GB/T 45654-2025 风险分类

接上网关：攻击被拦住，数据不出门

JJDGuard · AI 网关

员工 · 业务系统

智能体

JJDGuard 网关 本地判定模型

攻击识别 | 敏感数据脱敏 | 按人授权与额度 | 全量留证

DeepSeek

通义千问

豆包

企业私有模型

同一批攻击样本，接入前后

接入前

接入后

被攻破率大幅降低

每一次调用都留证

谁、何时、哪个模型、命中什么、花了多少



员工电脑上的 AI 助手，一个不漏登记到人

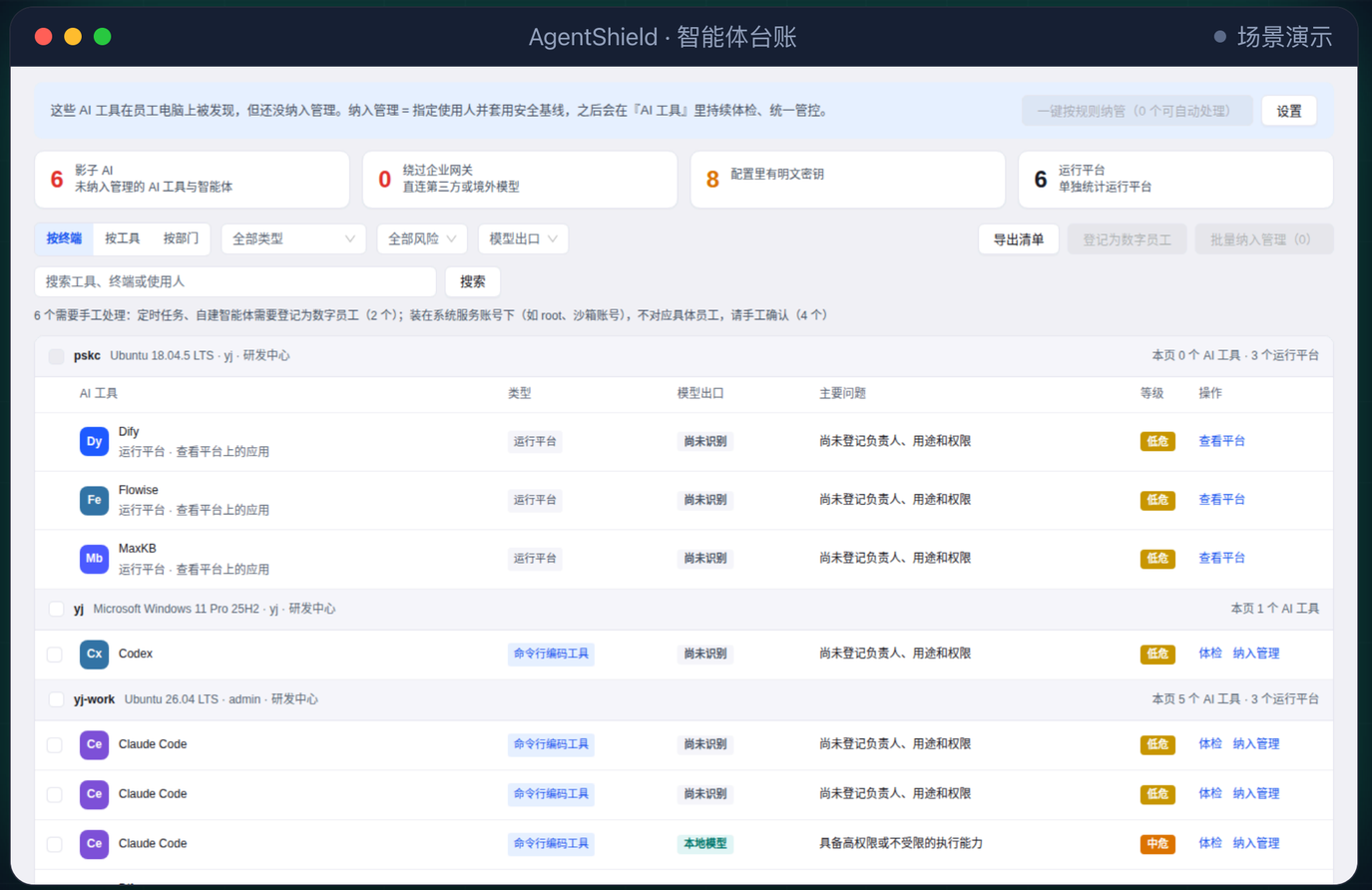
AgentShield · 智能体台账

员工电脑

Claude Code	Codex
Cursor	通义灵码
MCP 插件	技能包

AgentShield 台账 · 登记到责任人

- 影 未纳管的影子 AI
- 钥 配置里的明文密钥
- 助 编程助手登记到人
- 拦 钩子拦截：读取 ~/.ssh 私钥



下一步：像管员工一样管智能体

规划

● 鉴盾 · AI 体检

一次体检 → 常态化体检

变更自动复测
模型、提示词、知识库一变就重测

上线门禁
体检不过，智能体不上岗

测完自动出策略
发现的问题直接变成网关规则

● JJDGuard · AI 网关

模型网关 → 智能体网关

凭证自动轮换
每个智能体一把短期密钥，到期自动换

零改造接入
自动发现 AI 调用，自动下发配置

工具调用管控
高危动作走审批，异常消耗自动熔断

● AgentShield · 智能体台账

台账 → 数字员工档案

对接人事与 OA
入转调离，自动同步到名下智能体

全场景覆盖
终端、工作流平台、服务器上的智能体

行为基线
异常自动告警，一键隔离

汇成一件事 入职 → 发证 → 在岗 → 转岗 → 离职

让每一个智能体持证上岗，不越权、不干坏事

接下来，看真实产品

现场演示

01

鉴盾 · AI 体检

看一次完整的智能体体检

攻击轨迹 → 裁判判定 → 报告

02

JJDGuard · AI 网关

现场发一条带身份证号的消息

识别 → 脱敏 → 拦截 → 留证

03

AgentShield · 智能体台账

看员工电脑上的 AI 工具

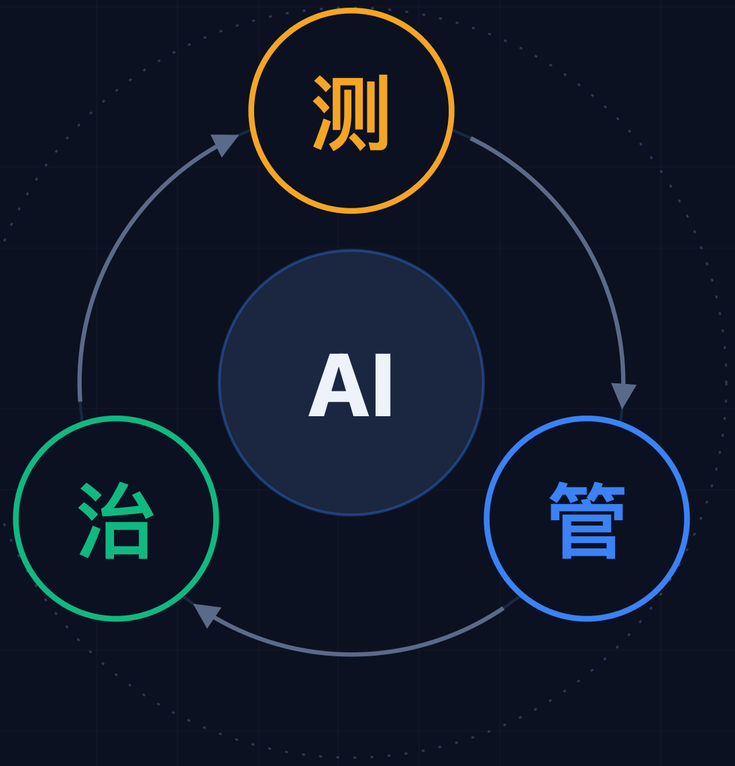
发现 → 登记到人 → 钩子拦截

演示环境，数据均为测试数据

谢谢

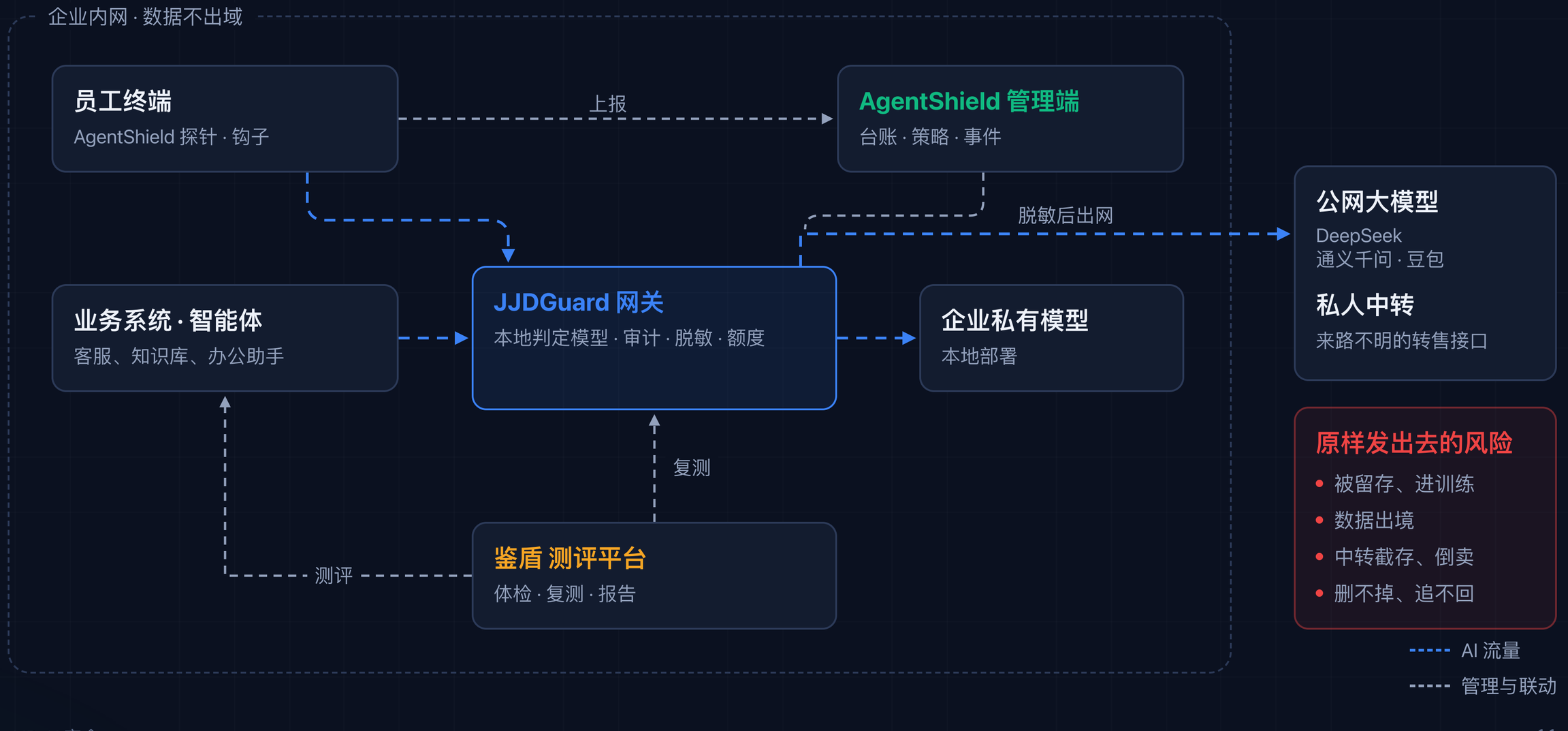
让企业敢把 AI 放进生产

微信扫码联系我们



小登队 · JJD AI 安全

附：部署架构


企业内网 · 数据不出域
员工终端
AgentShield 探针 · 钩子
上报
AgentShield 管理端
台账 · 策略 · 事件
脱敏后出网
公网大模型
DeepSeek
通义千问 · 豆包
私人中转
来路不明的转售接口
业务系统 · 智能体
客服、知识库、办公助手
JJDGuard 网关
本地判定模型 · 审计 · 脱敏 · 额度
企业私有模型
本地部署
复测
测评
鉴盾 测评平台
体检 · 复测 · 报告
原样发出去的风险
被留存、进训练
数据出境
中转截存、倒卖
删不掉、追不回
AI 流量
管理与联动