— 从POC到生产上线

企业AI应用运行时安全控制

云安全联盟 →

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

问题定义

01

控制面重构：AI应用上线的关键

企业AI应用上线的关键变化，不是模型能力更强，而是“后果”被放大。

PoC阶段 → **灰度/试运行** → **生产上线**

- 验证能力
- 人工看护、失败成本低

- 验证边界
- 引入策略、审批、日志

- 控制后果
- 每个访问和动作可仲裁

模型安全 ≠ 生产安全

AI应用 =“可被诱导、可误判、可越权的业务执行主体”

运行时访问控制 →把AI的不确定性限制在企业可接受的业务边界内。

02

PoC到生产：风险从“说错话”到“做错事”

PoC常见状态

- 离线样例
- 人工复制粘贴
- 只读知识库
- 少量试点账号

后果放大 →

生产真实状态

- 接入企业身份
- 读写业务系统
- 自动调用工具
- 跨部门流程闭环

提示注入　外部内容把助手推向错误动作

过度代理　工具能力大于任务需要

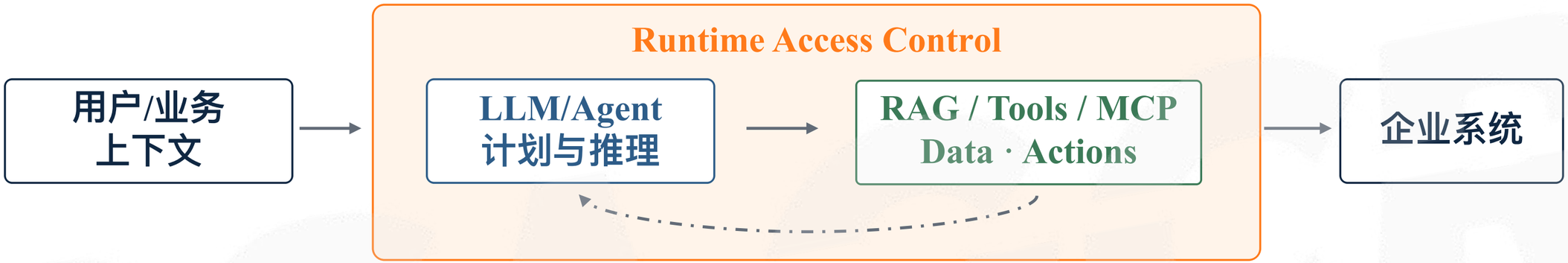
混淆代理　低权限上下文借高权限工具越权

审计断裂　知道“做了什么”，不知道“为什么被允许”

RAG 时代主要风险是“看见不该看的”；Agent 时代主要风险是“做了不该做的”。

03

AI Agent的运行时访问控制执行平面



控制什么?
Prompt / Context / Data / Tool / Action / Output

不做什么?
不试图控制模型思想；不靠系统提示词当安全边界

证明什么?
每次允许 / 拒绝 / 脱敏 / 审批都有证据链

模型可以计划，但访问和动作必须由控制平面仲裁。

02

架构展开

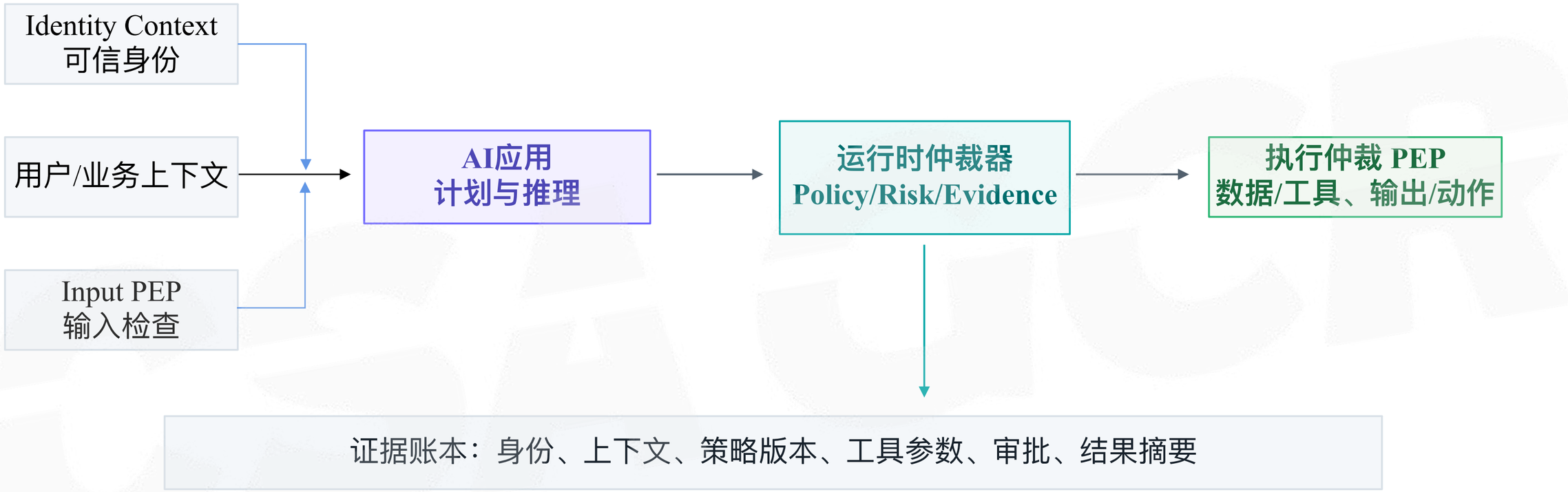
04

运行时访问控制执行平面的参考架构

最小权限　责任分离　完全仲裁

规范模型行为，而不是约束模型思想

运行时安全控制模型：把AI应用放进访问控制闭环



关键改造：不让模型直接拥有系统权限，而是让每次动作请求都经过运行时仲裁。

06

五类访问必须完全仲裁

Prompt Access

- 用户自然语言是否允许进入模型上下文

Context Access

- 历史状态、Memory，Session是否可用

Data Access

- Chunk, document, record, field是否可见

Tool Access

- 工具是否可见，参数是否可用

Action Access

- Send, write, delete, export等动作是否可执行

访问控制不是只在登录时发生；AI应用风险发生在多轮上下文、检索结果、工具调用和外部动作之间。

07

访问控制的基本原则

1 最小权限

只授予完成当前任务所需的最小动作、最小数据、最短时间。

2 责任分离

规划、审批、执行、审计不由同一主体完全掌控。

3 完全仲裁

每一次访问每一个对象都被检查，不能绕过Reference Monitor。

安全原则	系统设计	AI应用落地
Least Privilege	权限是否过大?	工具、数据、动作是否按任务收敛?
Separation of Responsibility	单独完成高风险动作?	模型、用户、审批人、执行器如何拆分?
Complete Mediation	所有访问都经过检查?	是否每次工具调用都被仲裁和留痕?

运行时架构：控制面、执行面、证据面三层闭环

把可变的业务策略和可变的模型行为，压到稳定的运行时仲裁路径里。

原则一：最小权限 = 动态能力令牌，而不是固定大权限账号

- **用户身份**：谁发起任务?
- **任务意图**：要完成什么?
- **数据范围**：只读哪部分?
- **工具动作**：只允许哪些操作?
- **有效时间**：何时失效?

→ **Capability Grant**
任务级临时授权

→ **Tool Call**
被约束的调用

生产红线

- 禁止全局读写
- 禁止长期Token
- 禁止模型直连系统
- 禁止默认继承用户全部权限

授权粒度建议： $user \times agent \times session \times task \times object \times action \times environment$

最小权限的策略对象：把“能做什么”写成机器可判定条件

Subject	谁	user、agent、service account、审批人
Object	访问对象	知识库文档、客户记录、工单、报表、配置项
Action	动作	read、summarize、export、write、send、approve
Context	上下文	任务、会话、数据等级、时间、网络、风险分
Effect	结果	allow / deny / mask / require-approval / read-only

典型规则：销售助手可读取本人负责客户的摘要字段；
导出、批量发送、跨区域客户明细必须升级审批。

策略不是文档；策略应成为运行时可执行的合同。

数据访问控制：RAG不是权限系统，Metadata才是边界

如果RAG支持Metadata

Pre-filter + Post-filter 双层过滤，避免未授权候选进入 prompt 或日志。

如果RAG不支持Metadata

Retriever Proxy + 外部 Metadata Store；必要时按索引 / namespace 隔离。

如果访问控制系统看不到chunk

退化为output guard, 不成为RAG Access Control。

Unauthorized chunk never enters prompt. No metadata, no automatic use.

工具治理：统一工具入口不是便利功能，而是集中授权

工具可见性控制

不同用户、任务、风险上下文下，LLM看到不同工具。

参数级授权

不是“能不能调用工具”，而是“能不能用这些参数访问这些字段”。

结果脱敏与审批

工具结果回到访问控制；外发、写入、导出、删除必须受审批控制。

LLM 不拿全量工具清单；Agent不持有真实业务系统凭证。

原则二：责任分离 = 不让一个主体同时计划、执行、证明自己正确

职责	禁止组合
AI Planner	不得直接持有写权限
Approver	不得审批自己发起的高风险动作
Executor	不得修改策略或证据
Auditor	只读，不参与业务执行

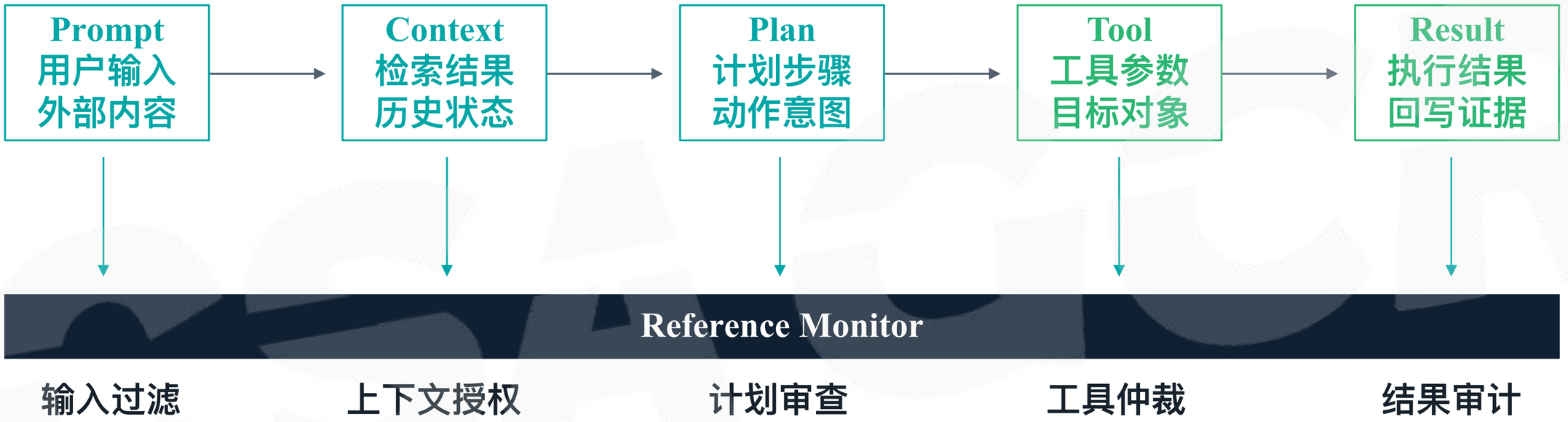
模型可以建议，不能独断；人可以批准，不能绕过；工具可以执行，不能自授。

按动作风险分层：哪些可以自动，哪些必须分权

L0 只读问答	允许自动；记录上下文
L1 摘要/改写	允许自动；敏感信息脱敏
L2 检索/分析	按数据等级授权；限制导出
L3 写入草稿	可生成不可发送；需用户确认
L4 业务提交	双人审批或流程审批
L5 配置变更/财务/外发	默认拒绝；Break-glass留痕

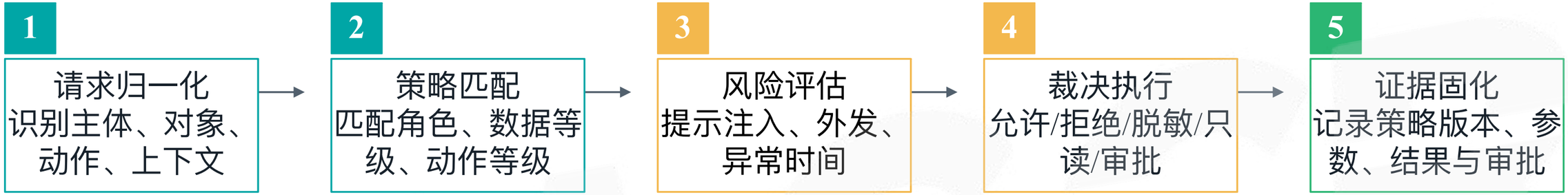
分层不是为了阻止AI，而是为了让AI应用在不同后果等级下使用不同控制强度。

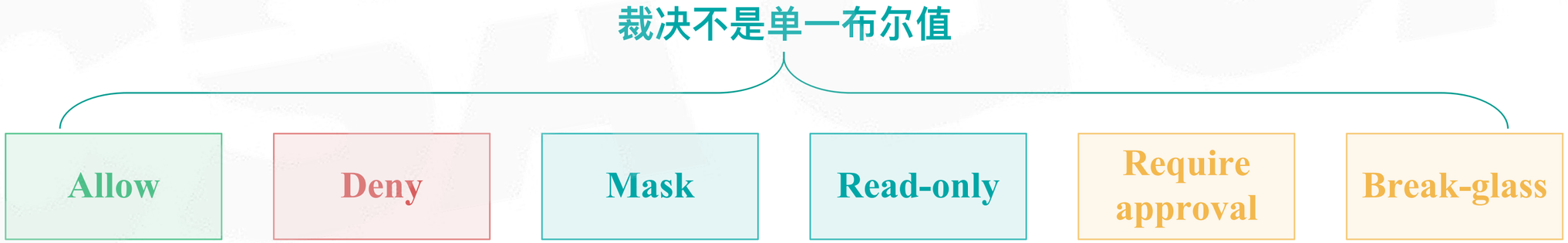
原则三：完全仲裁 = 每次访问每个对象都经过Reference Monitor



不能只在登录时鉴权；AI应用的风险发生在多轮上下文与工具调用之间。

运行时决策流：每一次工具调用都要产生一个可解释裁决





上线要求：任何裁决都应能回答“谁、因为什么规则、在什么上下文、允许/拒绝了什么”。

一次请求的运行时控制流程

Fail Closed: 关键依赖失败是默认拒绝或降级。

证据与审计：生产上线后，安全控制必须可证明

Intent 用户意图与任务边界

Context 检索内容、外部输入、会话状态

Policy 命中的策略ID、版本、例外

Decision 裁决、风险分、审批记录

Execution 工具、参数、目标对象、结果

Outcome 业务变化、回滚点、告警

审计问题必须从“谁调用了API？”升级为：

“为什么这个AI应用在当时被允许这样调用这个对象？”

工赋矿网
GONGFULIWANG

CSA GCR cloud security alliance®
GREATER CHINA REGION

03

场景验证

企业场景：同一控制面，覆盖不同上线故事

知识库助手

从“能回答”到“按权限回答”

- 按身份/部门/项目过滤检索
- 结果脱敏与引用回溯
- 导出行为升级控制

数据访问

BI/CRM助手

从“会分析”到“不会越权导出”

- 行级/列级权限查询
- SQL/DSL由网关重写校验
- 批量导出与外发受控

数据分析

IT/安全运维助手

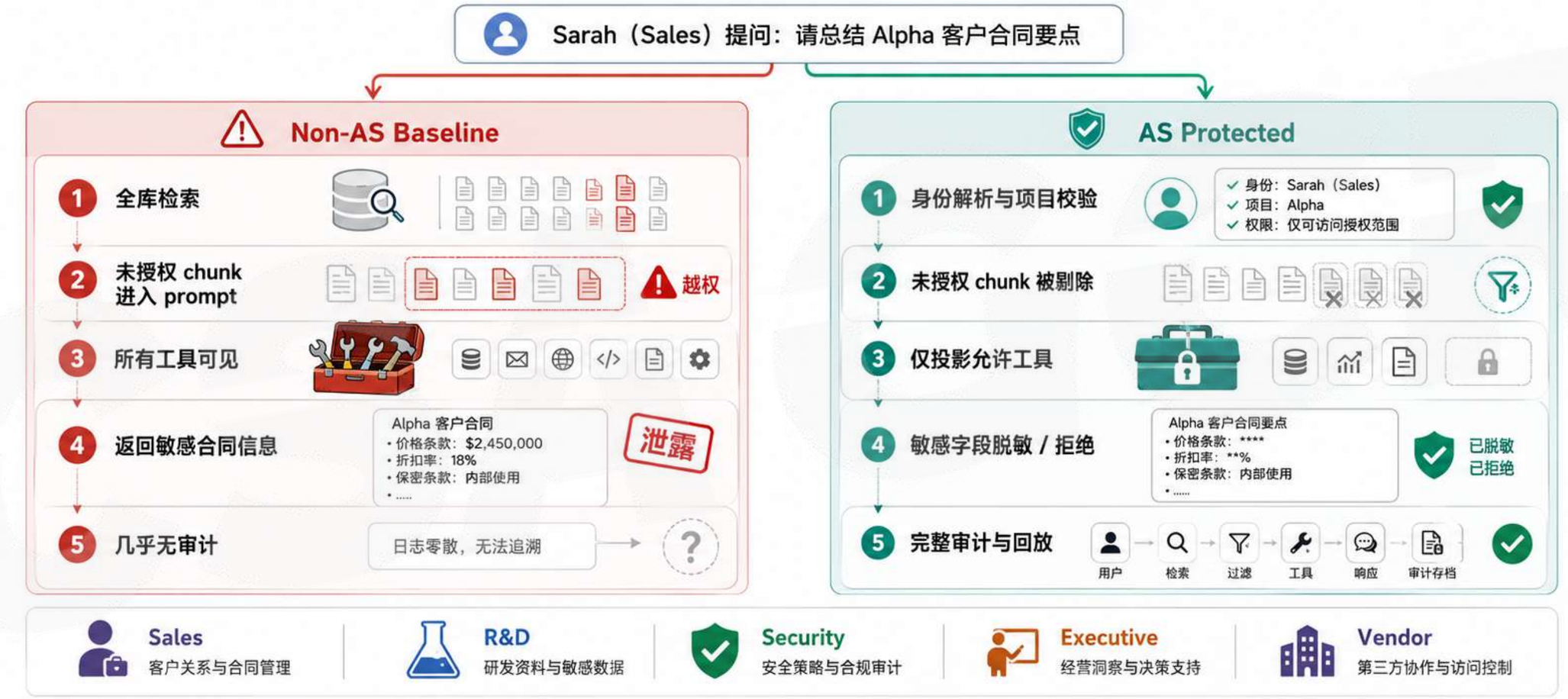
从“会处置”到“可控处置”

- 事件相关资产临时授权
- 高危动作审批/应急授权
- 变更可回滚且留痕

高权限动作

三类故事的共同点：模型不是权限边界，运行时仲裁器才是权限边界。

同一个问题，不同的运行时控制结果



让AI从“能回答”走向“受控的回答与行动”

04

上线方法

从PoC到生产的准入门槛：把上线拆成四道门

明确用户、数据、动作、失败后果

策略对象化；例外有Owner和过期时间

红队样例、越权测试、审批路径演练

指标、告警、复盘、策略迭代机制

准入不是一次评审，而是一组可自动检查、可回归测试、可追责的上线条件。

22

访问控制的验证矩阵

类别	必须覆盖的测试点
RAG/Data	Allow, Deny, Project Isolation, Clearance, Metadata missing, Semantic Leakage
Tool/MCP	Projection, Not visible deny, Field partial allow, Result sanitizer, Approval
Bypass/Fail-open	Raw disabled, Body identity ignored, Policy down deny, Audit failure block
Evidence	每个case输出audit_id, trace_id, allowed/denied chunks, visible/hidden tools

验证标准不是“回答成功”，而是“允许、拒绝、脱敏、审批、审计都可被机器化证明”。

指标与路线图：安全控制要能度量，也要能启动

Coverage	关键工具调用仲裁覆盖率	≥ 99%
Least Privilege	任务级授权比例	≥ 90%
SoD	高风险动作双控比例	100%
Evidence	完整证据链比例	≥ 99%

30天	边界与样板	身份、知识库、基础策略
60天	运行时与试点	Tool/MCP、审批、测试回归
90天	生产准入	生产准入、SIEM/GRC，报表

路线图不是从“更多功能”开始，而是从“安全不变量可验证”开始。

老生常谈：AI应用的生产上线条件

最小权限 拿到的是当前任务所需的临时能力，而不是系统大权限。

责任分离 不能单独完成高风险动作；计划、审批、执行、审计被拆开。

完全仲裁 每次访问数据和工具都经过不可绕过的运行时裁决。

PoC证明AI会做事；生产控制证明AI只能在被授权的边界内做事。

约束模型思想 < 控制应用行为

工赋矿网
GONGFULIWANG

CSA GCR cloud security alliance
GREATER CHINA REGION ®

感谢您的聆听和学习”

云安全联盟
大中华区 →