Tencent 腾讯 CSIG 云与智慧产业事业群

WorkBuddy 企业安全方案

——部署架构、端侧安全、业务数据安全、内容安全与安全建议

腾讯云 架构师 刘晓炜

2026.07.13

智能体时代到来，AI 正从辅助工具走向任务执行与组织协同

AI从工具走向智能体

AI从提效走向生产力重构

AI从辅助走向自主规划和任务执行

不仅是使用AI，而是**指挥和管理一群AI数字员工**

当 AI 开始自己干活，自己修复，自己交付结果，人类的工作方式，公司结构，甚至商业模式将会被重构，AI浪潮不是未来，是已经到来!

腾讯企业级智能体解决方案架构图

Tencent 腾讯 | 腾讯云

Agent Suite

- 腾讯文档
- 腾讯网盘
- 腾讯乐享

垂类智能体

- 研发智能体 (CodeBuddy)
- 设计智能体 (Ardot)
- 创意智能体 (Miora)

WorkBuddy · 企业 AI 工作台

超级个体 (个人提效)
使用 WorkBuddy 办公或指挥同事完成任务

超级团队 (组织提效)
通过 Teams 组织协同人或数字员工协同完成任务

沉淀企业资产

数字员工可由人或组织指挥

智能体构建 · WorkBuddy Build

7*24 小时在线的数字员工

专家	前端开发	项目经理	营销顾问	招聘经理
专家团	一人公司 (OPC)	软件开发团队	内容创作团队	

赋能数字员工

智能体与资产治理 · WorkBuddy Admin

企业 AI 资产

- **企业知识** 组织记忆
- **Skills** 技能封装
- **专家** 经验复制
- **连接器** 系统接入
- **大模型** 模型能力底座

01 WorkBuddy 企业版部署架构

TENCENT CLOUD

SaaS 企业版部署架构图

架构安全

- **租户数据隔离**
- **MCP 业务数据不出本地**
- **模型可只做推理**
- 操作日志支持审计

依托天御、信安、腾讯合规平台三重体系，全链路严把内容合规关口

VPC 企业版部署架构图

Tencent 腾讯 | 腾讯云

客户接入层

用户接入终端

Mac | Windows | Linux | Web | 小程序 | API

WorkBuddy 客户端

- 默认权限·安全沙箱
- 完全放开·无需确认

安全机制（默认权限·安全沙箱模式专属）

· 沙箱命令仅在沙箱级别执行
· 无法修改工作目录外的文件
· 危险指令自动拦截二次确认

完全放开：操作无需确认直接执行

MCP 调度 | 第三方 MCP

MCP 工具调用均在客户端本地执行

客户 PVC

域名 SSL

VIP 私有链接

独享 VPC

腾讯云 CLB

网关层

鉴权 路由分发

后端服务

AI 核心服务
对话管理·AgentCore·Agent调度
推理链路·后台Agent·Prompt管理·Team Agent

业务服务
登录认证·企业策略·组织管理·用户中心
管理控制台·知识库管理·开放平台

Agent 扩展服务
记忆查询·对话清洗
工具注册·记忆提取

平台服务
控制台·审计日志

向量服务
向量检索·文本向量化

MCP 执行不在后端服务

中间件层

数据存储：MySQL | Redis | Milvus | ClickHouse

对象存储：COS | 数据备份

可观测：Grafana | Prometheus | Loki | Tempo

基础设施：Kubernetes Helm Docker
amd64 / arm64 多架构支持
OTel | cAdvisor

模型接入层

Model-Gateway-1

Model-Gateway-2

Model-Gateway-3

Model-Gateway-4

模型推理服务层

数据隐私保障
仅进行模型推理
不记录任何客户数据
安全管控

模型推理

GLM-5.1
Kimi 2.6
Minimax 2.7
DeepSeek-V3/R1
混元大模型
自定义模型

架构安全：

- VPC 级别数据隔离
- MCP 业务数据不出本地
- 模型可只做推理
- 操作日志支持审计
- 依托天御、信安、腾讯合规平台三重体系，全链路严把内容合规关口

私有化企业版部署架构图

架构安全：

- 完全私有化，数据不出客户环境
- MCP 业务数据不出本地
- 操作日志支持审计

02 WorkBuddy
端侧安全

TENCENT CLOUD

端侧典型威胁场景

威胁 1：非预期文件操作

痛点：模型被 Prompt 注入或推理出错时，产生 rm -rf、覆写用户文档、越权访问凭据文件等动作，影响范围超出当前工作目录

需求：安全兜底：即便模型出错，破坏半径被限制在可控范围。

“谁”来划定破坏边界 + 统一运行环境？ → 安全沙箱

威胁 2：命令执行失控

痛点：AI 在终端调用高危命令（磁盘格式化、进程终止、注册表修改、停止安全服务等），或串联多条命令实现意料之外的破坏链

需求：看懂行为：识别 AI 每一步工具调用的真实意图，对已知威胁做执行前判定

如何对每一条执行命令进行安全行为判定 → 行为检测能力

威胁 3：Skill 注入 / 供应链投毒

痛点：被污染的 Skill 诱导 Agent 做非预期动作，绕过正常业务逻辑，Agent 自动安装 / 加载第三方 Skill 场景被投毒，通过生态渠道注入恶意代码

需求：针对危险 Skill 进行安全检测，杜绝危险 skill 执行

如何防止 Skill 注入和投毒 → 生态安全检测能力

WorkBuddy 端侧安全方案说明

Tencent 腾讯 | 腾讯云

安全沙箱
- 文件隔离
- 行为审计
- 运行环境一致化

行为检测能力
- 本地 SDK
- 云端判断

生态检测能力
- Skill 检测

自主能力（长期规划）
- 执行链路全可控
- 自动化处置
- 用户低介入

安全沙箱
3 类
安全措施

文件隔离
把 AI 自主产生的文件读写、命令执行、网络外连收敛在可控边界内——工作目录以外的文件不会被改动

行为审计
运行时 Hook 采集文件 / 进程 / 命令事件，输出 JSONL，在本地和服务端留存

运行环境一致化

行为检测能力
2 类 工具 +
腾讯情报底座

本地 SDK
嵌入 Agent Runtime，命令内容的快速检测在本地完成，零延迟、离线可用

云端判定

Skill 检测
4 类 手段 +
腾讯情报底座

病毒查杀
基于腾讯威胁情报底座的反病毒引擎，对 Skill 包内文件做 Hash / 签名 / 静态特征命中

算法识别

落地形态
在 Skill 加载前完成检测，识别为高风险时阻断加载，并产出结构化告警事件

沙箱模拟

WorkBuddy 端侧：安全沙箱管控

安全管控

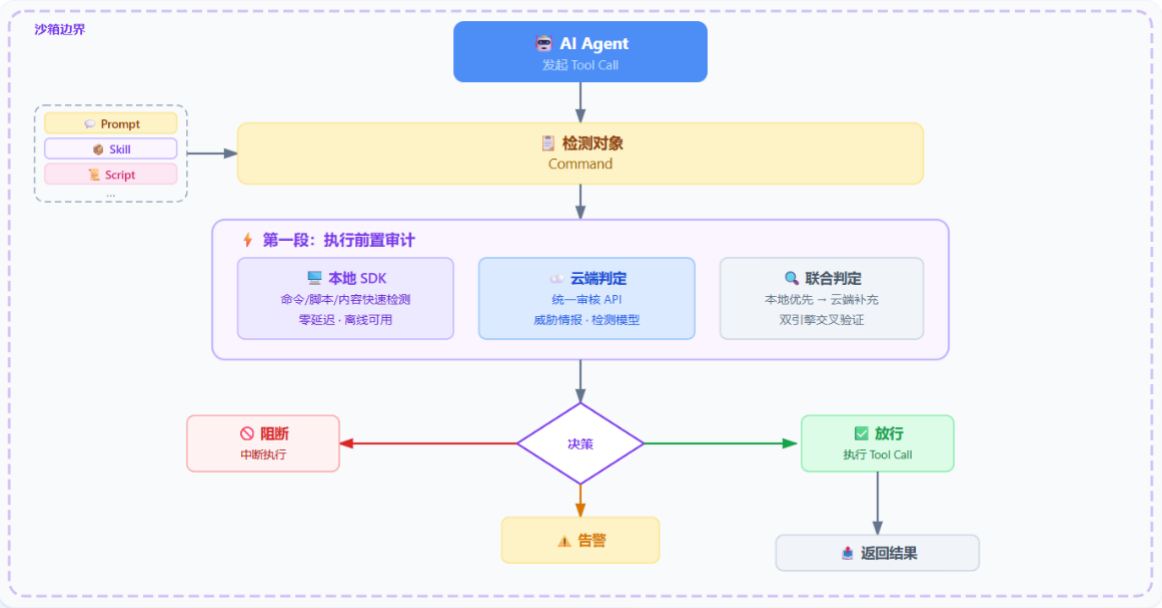
把 AI 自主产生的文件读写、命令执行收敛在可控边界内——工作目录以外的文件不会被改动，关键行为全部写入 JSONL 结构化审计日志。

运行环境一致化

沙箱内嵌 Python / Node.js + npm / Git / 等完整工具链，让 AI 产物在任何用户本机都能跑出一致结果。

macOS 运行时沙箱	*基于 sandbox-exec，文件 / 进程隔离 + 访问审计*
Windows 运行时沙箱	*增强型 Hook，ntdll Hook + 子进程传播 + 与主流安全软件共存；文件 / 进程隔离 + JSONL 审计*

WorkBuddy 端侧安全：行为检测（5.0 版本接入）



本地 SDK： 在沙箱边界之内，看懂每一步 Tool Call 的真实意图，对已知威胁做执行前拦截。沙箱管"能不能动"，行为检测管"该不该动"。

架构组成

本地 SDK（5.0 版本）：

嵌入 Agent Runtime，命令的快速检测在本地完成，零延迟。

云端判定（5.0 版本）：

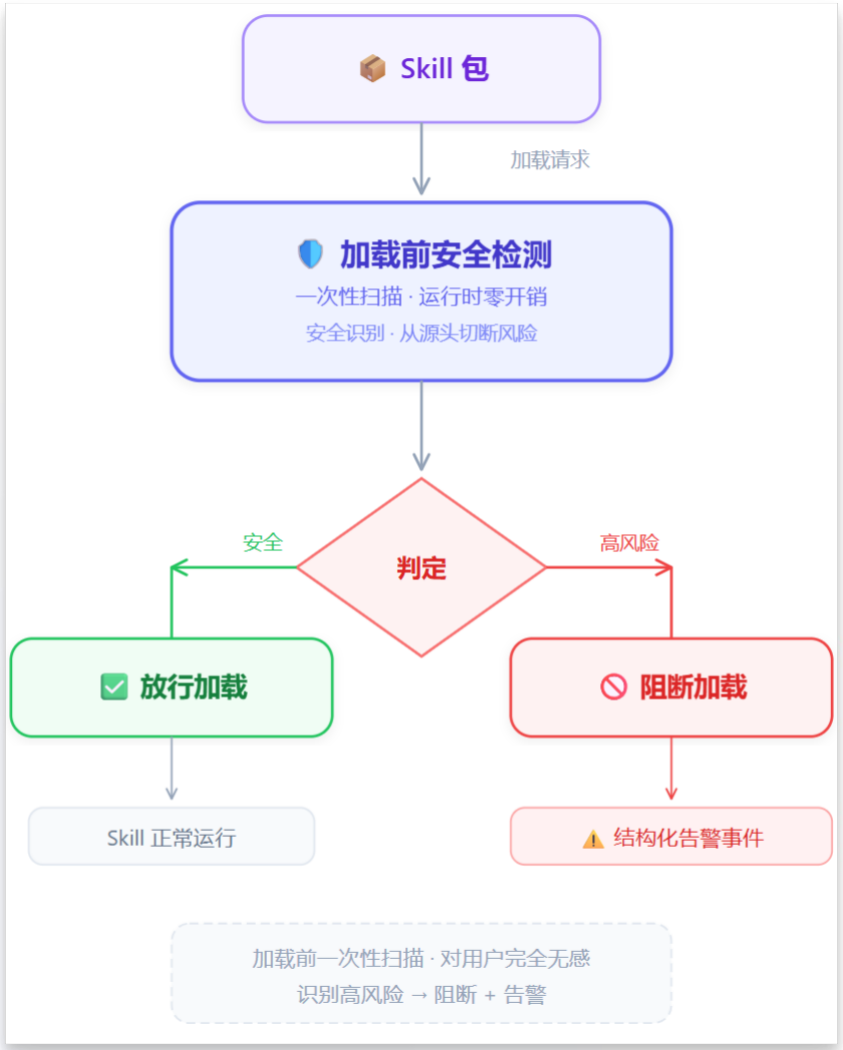
统一审核 API，动作区分 content / script；云端汇聚规则、检测模型与腾讯威胁情报（文件查杀、恶意 IP / Domain、供应链包依赖）等。

检测对象与时机

对象： Command（命令执行）。

决策： 放行 / 标记告警 / 阻断；阻断时可直接中断工具。

WorkBuddy 端侧安全：Skill 检测



腾讯优势

- **威胁情报中的病毒查杀能力**

  直接复用腾讯全生态共建共用的反病毒引擎与样本库，已知恶意 Skill / 包 / 文件**秒级命中**；情报库随腾讯安全运营持续更新，客户端**无需发版**。

- **科恩实验室的算法能力**

  科恩在二进制、漏洞、移动端、Web、AI 安全等方向有长期算法积累，把这些经验沉淀为面向 Skill 的恶意识别算法，覆盖单纯靠规则与情报无法识别的**"未知 Skill 风险"**。

WorkBuddy 端侧安全：规划中能力

Windows 沙箱演进至原生安全原语组合 规划中

迁移到 **Restricted Token + Job Objects** 等系统原生机制，稳定性与可维护性来自系统本身；与系统更新、其他安全软件共存更顺畅。**对外能力接口不变，升级对产品层与用户无感。**

自动化处置能力 长期规划

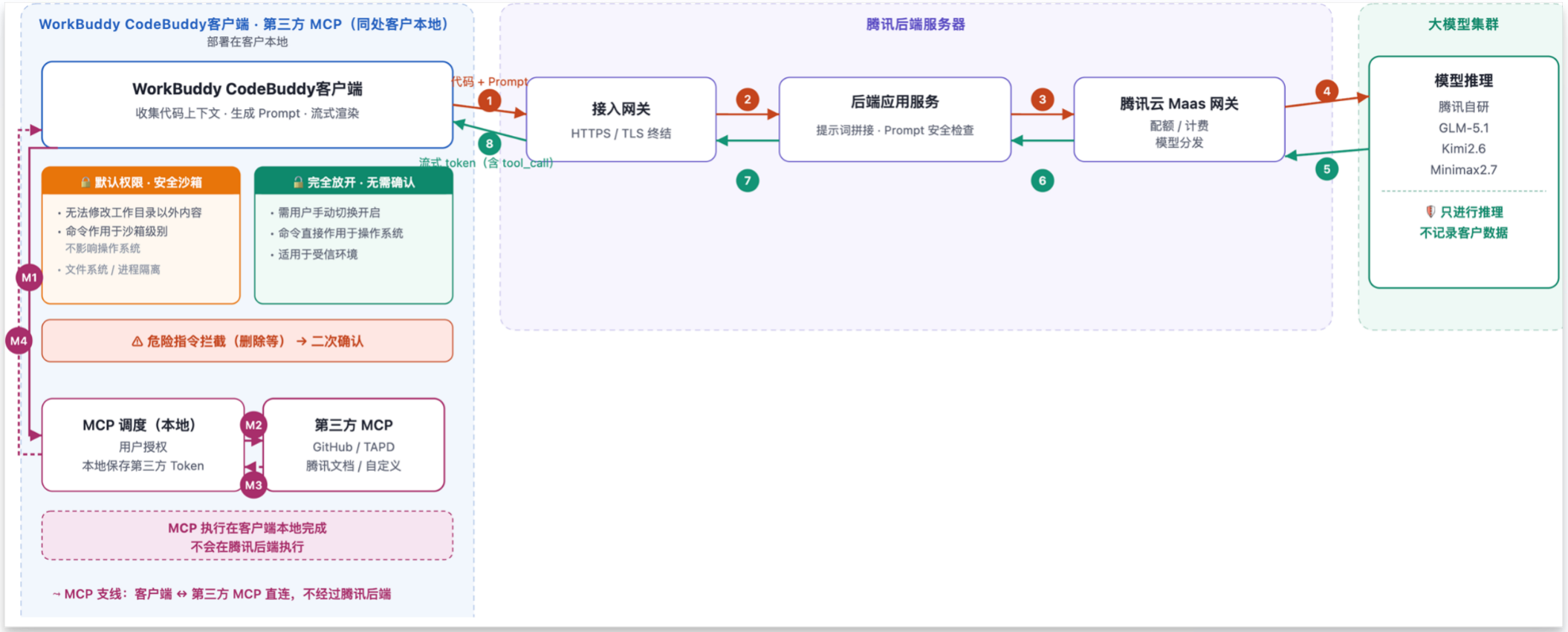
沙箱进出 / 放通 / 拦截 / 询问由系统自动决策；低风险场景静默放通，显著减少授权弹窗；**让 Agent 自动化真正跑起来。**

03 WorkBuddy 业务数据安全(MCP)

TENCENT CLOUD

WorkBuddy：业务数据安全

> MCP 业务数据交互在客户端直连完成，数据不经腾讯后端，保障企业数据安全。



04 WorkBuddy AIGC侧业务安全

TENCENT CLOUD

WorkBuddy：AIGC 侧安全

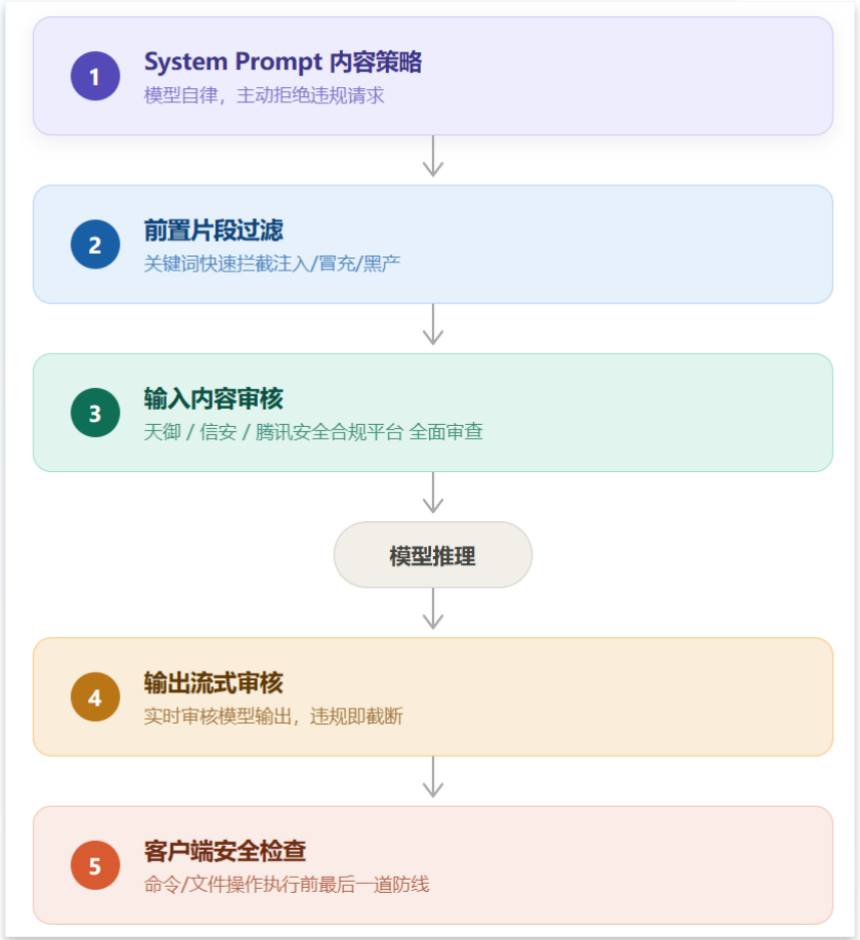
1 可必要数据进行推理、不训练

推理层只用数据运算、用完即删，
不留存、不训练 也不外泄用户问答数据

2 腾讯安全管控模型内容合规

依托天御、信安、腾讯合规平台三重体系，
全链路严把内容合规关口，
全链路守护数据合规无忧

WorkBuddy：AIGC 整体安全流程（大模型内容五道安全防线）



1 第一道防线

系统提示中内置安全策略，让模型自主拒绝违规请求

2 第二道防线

调用审核 API 前，毫秒级精确匹配已知攻击关键词拦截

3 第三道防线

将用户输入送天御 / 信安 / 腾讯安全合规平台全面审查，支持图文联合

4 第四道防线

模型流式生成时实时增量审核，发现违规立即取消推理并截断输出

5 第五道防线

Agent 执行终端命令或文件操作前，客户端本地做最后一道风险拦截

05 WorkBuddy 端侧操作安全建议

TENCENT CLOUD

5.1 使用决策：什么该用、什么不该用

核心原则：遵循 80 / 15 / 5 数据分级法则，确保业务效率与数据安全的完美平衡。

80%

直接使用

无个人信息的公开或通用数据，可直接用于日常业务处理。

- ✓ 公开的新闻稿件与宣传文案
- ✓ 无敏感数据的产品说明书
- ✓ 通用的技术文档与操作指南

15%

需脱敏使用

包含部分隐私或内部信息，必须经过匿名化或掩码处理后方可使用。

- 包含用户真实姓名的业务报表
- 带有手机号、邮箱的客户记录
- 带有真实系统IP地址的运行日志

5%

绝对不碰

涉及系统安全与商业机密的核心数据，严禁输入到任何外部工具中。

- ✕ 系统账号密码与访问凭证
- ✕ API Key、Token 与加密密钥
- ✕ 未公开的商业合同与财务数据

5.2 权限边界：给AI多大的操作权限

什么时候该降级

任务特征	建议模式	说明
任务目标明确，流程熟悉	Craft	风险可控，直接执行效率最高
涉及多文件批量操作，怕跑偏	Plan	先看方案，确认了再动手
只问问题，不需要操作	Ask	零风险，AI 碰不到任何文件
不确定该不该让 AI 动手	先 Ask → 再决定	不要在犹豫时直接给最高权限

关键场景：
处理重要文件前先备份 → 降到 Plan 预览改动 → 确认无误再执行

核心原则

拿不准的时候，退一档。权限这件事，够用就行，多了反而危险。

文件授权：别给 AI 整把钥匙

做法	风险 / 优势
✕ 授权整个桌面	个人照片、合同扫描件、银行回单全暴露
✕ 授权下载目录	临时文件、压缩包、敏感附件混杂
✓ 只授权临时目录	只放本次必需文件，任务结束移除

最小授权原则（四步法）：

1. 新建临时文件夹
2. 放入本次必需文件
3. 只授权该文件夹
4. 完成后移除授权

真实教训

小李让 AI 批量整理文件，图省事授权了整个桌面。AI 在执行过程中读到了桌面上一份尚未公开的报价单，并在生成的汇总报告中引用了其中的价格数据。虽然文件没离开电脑，但这份含敏感数据的报告如果被不小心转发，后果不堪设想。

给 AI 的权限，刚好够完成这件事就行。
多给一份，就多一份风险。

5.3 输出验证：AI的回答能信多少

建立可靠的验证机制：三步核实法

1. 核数据

验证底层信息的真实性与时效性

- 来源是否权威可靠
- 日期是否最新有效
- 人名/机构名无误
- 核心数字精确无造假

2. 核逻辑

审查推理过程的严密性

- 推理链路是否完整
- 因果关系是否成立
- 论证过程有无跳跃
- 前后观点避免自相矛盾

3. 核场景

评估实际应用的匹配度

- 是否符合实际业务需求
- 方案落地是否具备可行性
- 语境与受众是否适配
- 与最终目标是否对齐

核心原则：AI输出永远是「初稿」

请务必保持独立思考，警惕"AI幻觉"（一本正经地胡说八道）带来的潜在业务风险。

5.4 防线配置：五个入口一个不能松

01 账号安全

- 强制启用多因素认证(MFA)
- 严格实施最小权限原则(PoLP)
- 定期审计账号活跃度与权限

02 API Key保护

- 严禁在代码库中硬编码密钥
- 统一使用密钥管理服务(KMS)
- 设定严格的调用频次与IP白名单

03 Skill安装规范

- 仅限从受信任的官方渠道获取
- 安装前执行自动化安全与漏洞扫描
- 严格限制Skill的系统级访问权限

04 远程控制安全

- 全面采用零信任网络访问(ZTNA)
- 高危操作需二次授权与全程录屏
- 部署异常行为检测与实时阻断机制

05 环境隔离

- 生产、测试与开发环境严格物理/逻辑隔离
- 实施微服务间的网络微隔离策略
- 敏感数据全生命周期加密存储与传输

核心安全规范

构建坚不可摧的云端防线，遵循三大基本原则

责任共担

平台保护 ≠ 用户免责

WorkBuddy 提供底层基础设施的物理与网络安全保障，但用户必须对自身的数据、应用配置、身份凭证及访问控制负起最终责任。安全是双方共同构建的防线，缺一不可。

权限最小化

不确定时“退一档”使用

严格遵循“Need-to-know”原则。仅授予实体完成特定任务所需的最低权限级别。在权限边界模糊或需求不明确时，宁可降级授权，以最大程度降低因权限滥用或凭证泄露带来的潜在风险。

警惕幻觉

人工核实是不可逾越的底线

在引入AI辅助决策、自动化运维或智能代码生成时，必须时刻防范模型“幻觉”。任何涉及核心资产的关键操作与敏感变更，都必须经过严格的人工复核机制，确保系统的绝对可靠与可控。

感谢倾听