WorkBuddy: 企业AI工作台

安全方案

2026-06-02

腾讯云-云产品六部

Tencent Cloud

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

2025.01
DeepSeek R1 发布
Agentt 开始具备更强的思考能力

2025.02/03
Vibe Coding/Manus
智能体开始承接开放任务，自主拆解与推进
Agent 从回答问题走向完成任务

2025.05
Claude Code 发布
智能体进入代码执行与文件操作
Agent 从“会说”走向“会做”

2025.07 ★
CodeBuddy IDE 发布
全球首个全栈 AI 工程师
Agent 成为持续协作的开发伙伴

2026.1
Claude Code 推出 Cowork，OpenClaw 发布
多 Agent 协同、共享上下文与任务分工开始成型
Agent 从单体助手走向协作型数字同事

2026.2
OpenAI 提出 Harness Engineering
智能体竞争从“模型能力”转向“运行系统能力”

2026.3 ★
WorkBuddy 发布
企业AI工作台形态成型
Agent 从个人助手走向企业级工作台

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

腾讯云

01

WorkBuddy：产品版本与部署架构

产品版本、部署架构

Tencent Cloud

WorkBuddy 版本形态全览

SaaS · 免费

个人体验版

SaaS 公有云

🎯 初次体验 / 轻度个人用户

↑ 对话频率有限制

- AI对话问答
- 内置Skills
- 热门知识库

SaaS · 付费

个人专业版

SaaS 公有云

🎯 专业个人 / 自由职业者

↑ 含体验版全部能力

- 不限频对话
- 自定义Prompt
- MCP扩展
- 多模型协同
- 自定义Skills

SaaS · 企业

SaaS企业版

SaaS 公有云

🎯 中小企业 / 成长型团队

↑ 含专业版全部能力

- 企业知识库
- CloudAgents
- Teams协同
- SSO单点登录
- 安全审计

VPC · 专享

专有云企业版

VPC 专有云

🎯 大型企业 / 数据敏感场景

- VPC独立部署
- 数据不出内网
- SSL域名隔离
- 专属网站
- 专属logo

私有化 即将推出

私有化企业版

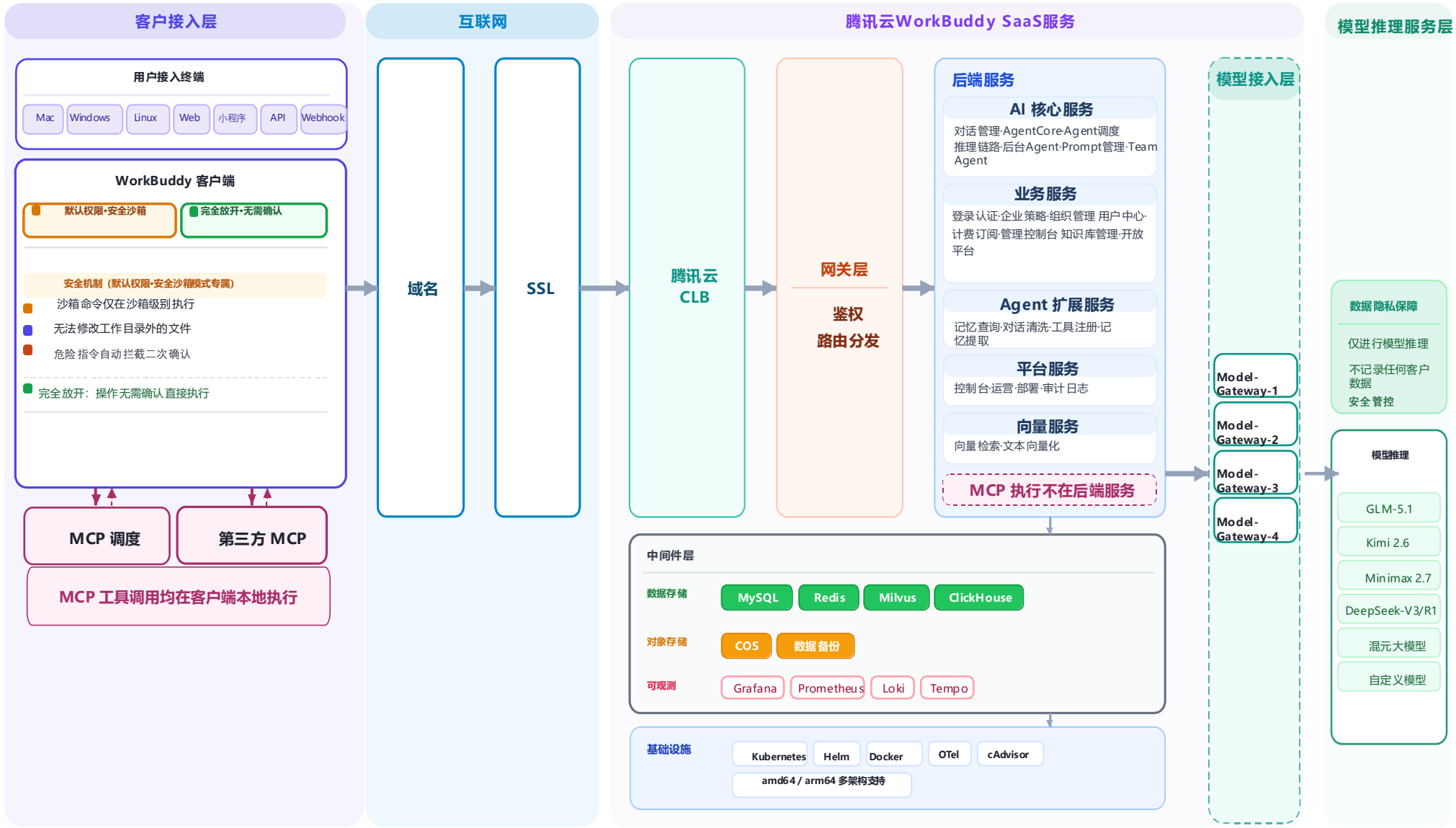
企业本地部署

🎯 政府 / 金融 / 超高安全

- 完全内网隔离
- 本地模型部署

提供五种版本形态，覆盖从个人到大型企业的全场景安全需求

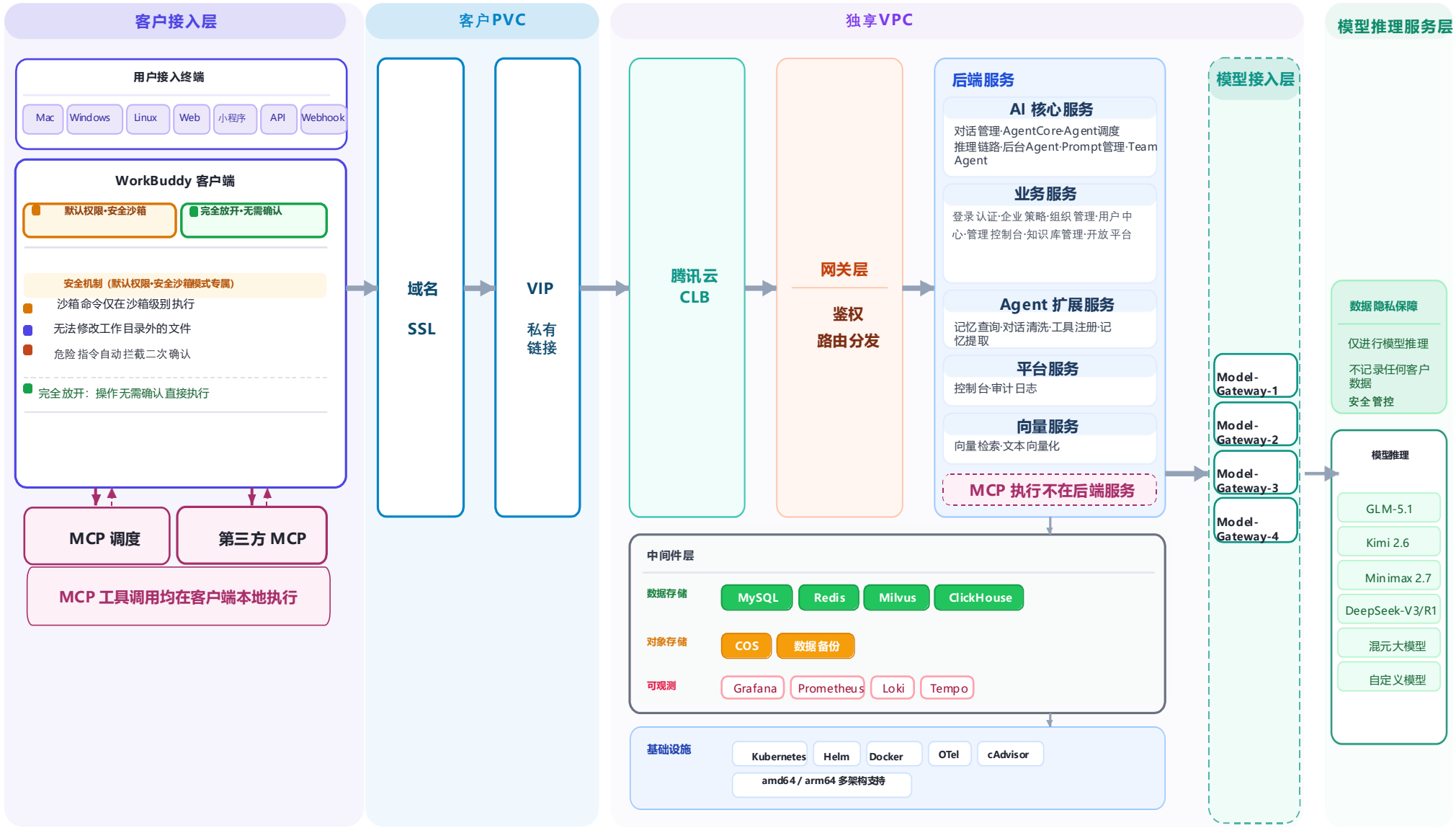
Saas企业版部署架构图



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

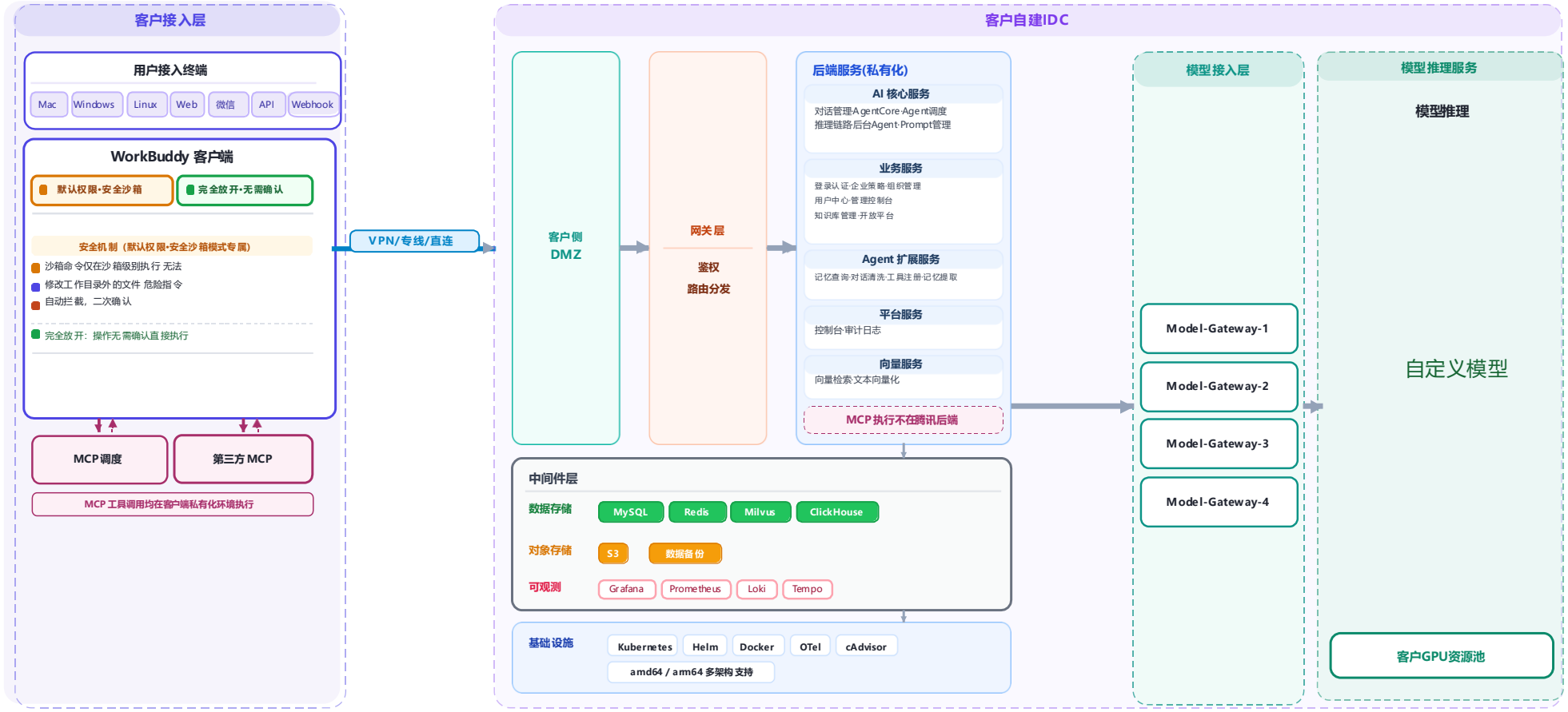
VPC企业版部署架构图



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

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ski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

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

运行环境一致化

Git Bash 提供工具链，进程为标准 Windows PE，受沙箱 Hook 直接覆盖

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

统一审核 API，云端汇聚规则，依赖腾讯腾讯全生态共用的安全情报底座。

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

腾讯科恩自研算法语义甄别 Skill 代码风险

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

远端沙箱仿真运行Skill，风险则拦截安装

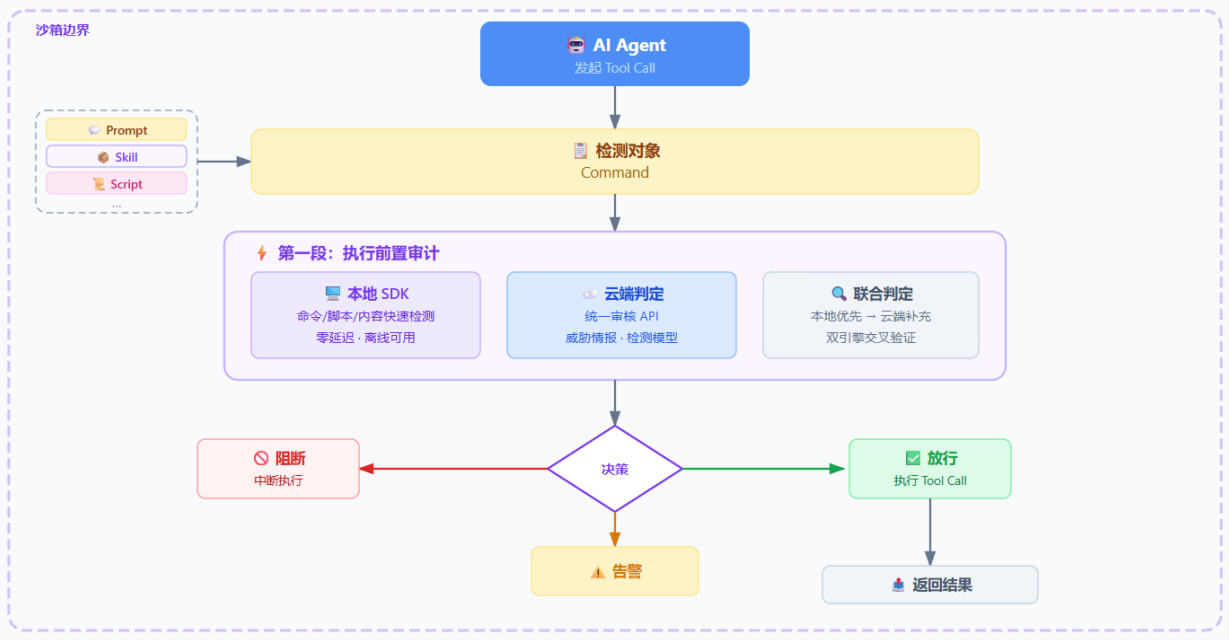
WorkBuddy端侧：安全沙箱管控（已接入）

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。

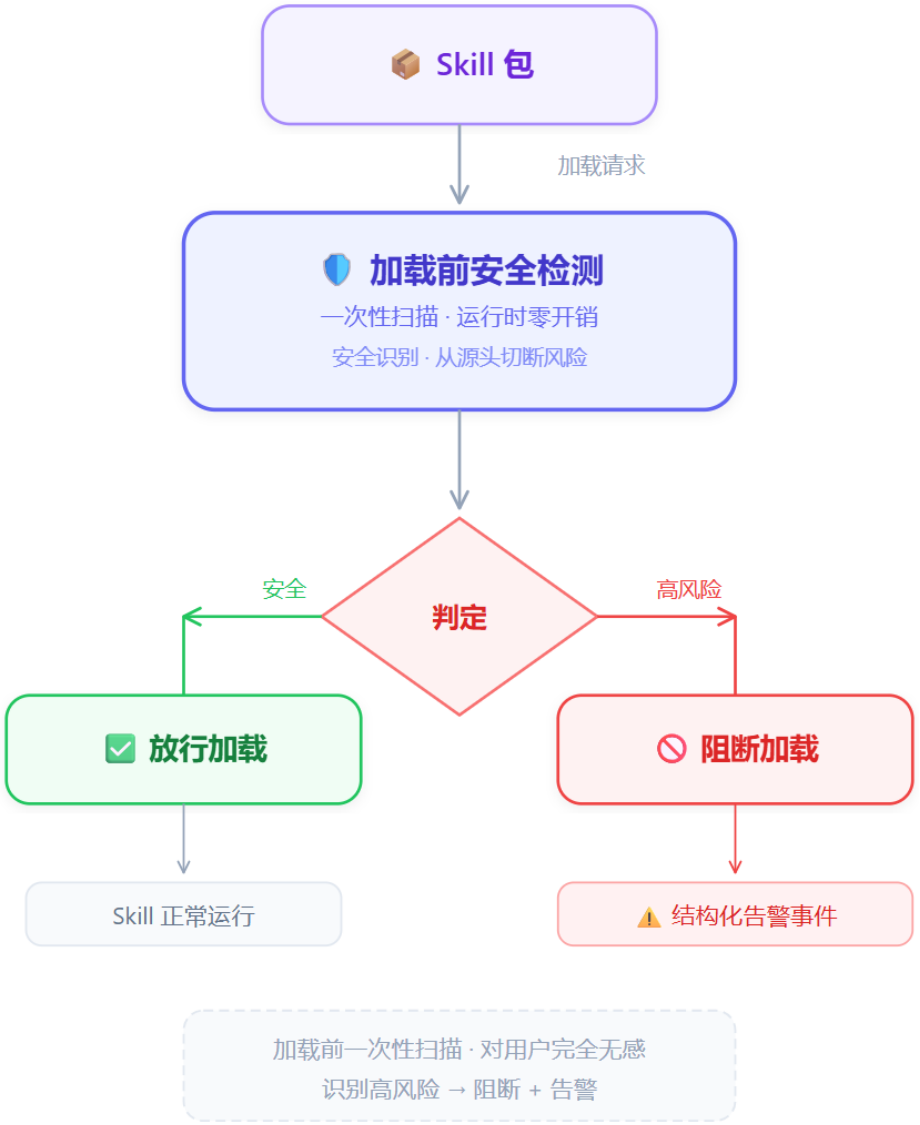
云端判定（5.0版本）： 统一审核 API，动作区分 content / script；云端汇聚规则、检测模型与腾讯威胁情报（文件查杀、恶意 IP / Domain、供应链包依赖）等。

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（已接入）



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。**

腾讯云

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。

腾讯云

04

WorkBuddy： AIGC侧数据安全

Tencent Cloud

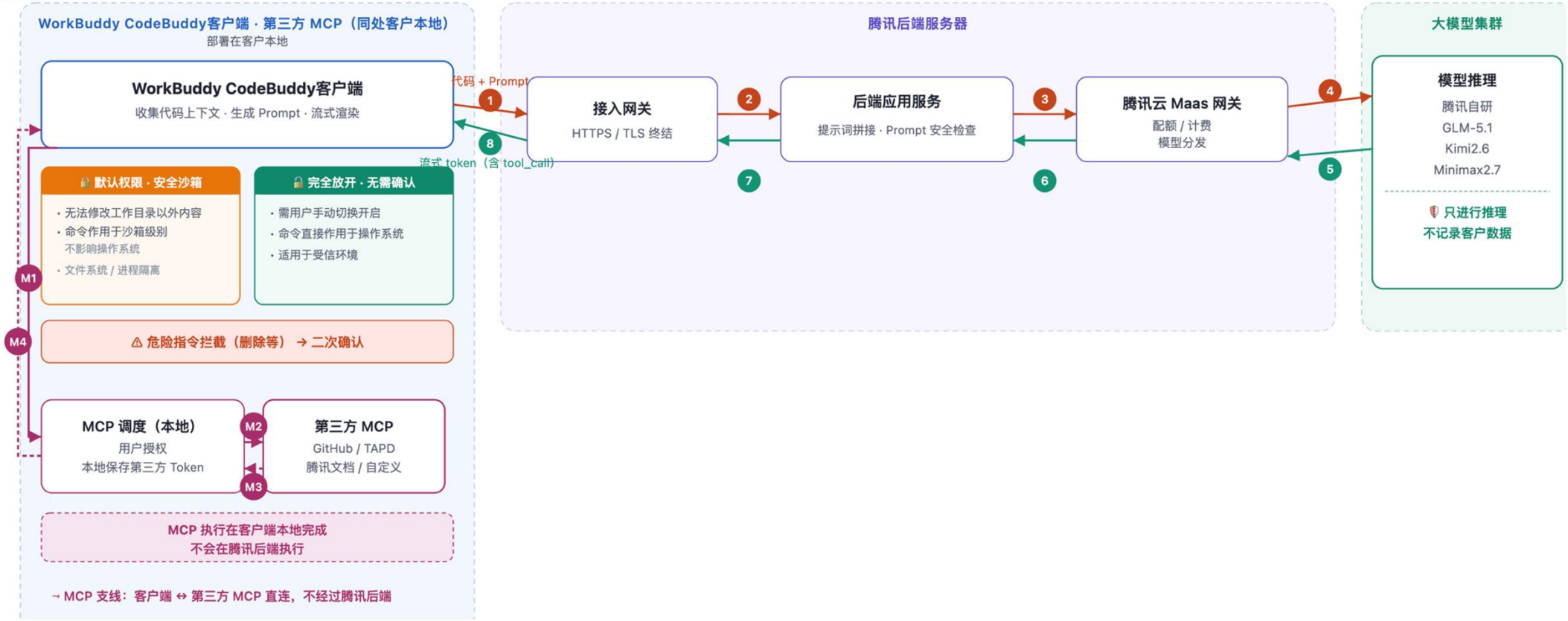
WorkBuddy：AIGC侧安全

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）

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

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

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

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

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

腾讯云

05

WorkBuddy： 端侧操作安全建议

Tencent Cloud

CloudAgent端侧数据安全建议

https://workbuddy-course.pages.woa.com/

《第五章·安全与合规》

THANKS

感谢您的聆听!

腾讯云