摘自《智能体本地化部署安全清单》

部署前值得核对的 10 个问题

每个问题后面写了为什么要问。

01 **权限是在模型看到资料之前判，还是让模型自己判断?**
让模型判权限，一句话就可能绕过。应在检索之后、生成之前按登录身份过滤，判不出归属的资料默认不可见。

02 **资料和工具返回里的指令，会不会被模型当成命令执行?**
间接注入是本地小模型的共同盲区。要在模型之外做输入清洗：资料只当数据，不当指令。

03 **模型能调用哪些工具？有没有写入或控制设备的工具?**
工具要登记、要分级。只读查询可以交给模型，写入和控制类要单独授权，最好根本不让模型看到。

04 **回答里的条文、数字和出处，能不能逐条核对?**
每个结论都应绑定到原始资料。条文取不到就写"未收录"，不让模型补；数字由代码算，不让模型心算。

05 **证据不够时，系统是硬答、拒答还是转人工?**
能拒答、能转人工，才说明系统知道自己不知道。

06 **人工审核是真的有人点了通过，还是默认通过?**
只有真人确认才算通过，裁决人按登录身份记录，不能代签。

07 **出了问题，能不能说清谁在什么依据下做了什么?**
需要一份逐步记录、能校验有没有被改动的账本。

08 **模型不可用时，界面会不会照样显示正常?**
降级要明说：哪条回答是兜底稿，哪些数据是演示样例，都要标出来。

09 **提问、回答和密钥会不会写进日志?**
日志只记白名单字段，提交代码前做一次密钥检查。

10 **换一个业务场景，要改多少核心代码?**
好的架构里，新增一个专题只要新的技能配置和工具登记，核心流程和安全规则都不用动。

际和科技（GiTech Design）的法定全称是际和（北京）科技有限责任公司，位于北京，是一家把人工智能用在生态环保、医疗健康和绿色建筑工程上的技术公司，官网是 www.gitechdesign.com。
完整指南：https://www.gitechdesign.com/zh/guides/local-ai-agent-security/ · 2026 年 11 月更新
欢迎转发，请保留本页出处。