M3 · 职业版 Agent · 第 1 讲

第 1 讲：Agent 架构与边界

先画审核层，再谈智能

阅读约 25 分钟

练习约 30 分钟

产出：五层架构图

本节目录

一个真实故障

本讲目标

一、五层架构

二、什么不该交给 Agent

三、三种架构怎么选

四、边界写在哪

五、30 分钟练习

六、验收清单

七、延伸阅读

一个真实故障

团队把“会议纪要助手”推广到全公司。第一周很顺，第二周开始出现：有人让它“顺便把纪要发到客户群”，它真的发了；有人上传了合同，它把条款写进了公开文档。

问题不在模型，而在架构里没有“边界”这一层。

本讲目标

- 画出 Agent 的五层架构；
- 判断哪些任务不该交给 Agent；
- 在单 Agent、工作流、多 Agent 之间做选择；
- 把边界写进系统提示与权限。

一、五层架构

层	作用	缺了会怎样
入口	接收任务、识别用户	谁都能触发
编排	拆步骤、决定下一步	一步到底，无法纠错
工具	读写文件、调接口	只会聊天
记忆	会话、长期偏好、知识库	每次从零开始
审核	人工确认、日志	出错无人兜底

画架构图时，先画审核层。没有审核层的 Agent，默认不能碰外部系统。

二、什么不该交给 Agent

四类任务先划出去：

1. **不可逆**：删除、付款、对外发送；
2. **高风险**：合同、医疗、法律结论；

3. **需要价值判断**：裁员名单、绩效评级；

4. **责任不清**：出了事没人能签字。

这些不是“不能做”，而是“必须有人确认”。确认节点要写进流程，不是靠提示词祈祷。

三、三种架构怎么选

架构	适合	代价
单 Agent	一个窄任务、步骤少	能力有限
工作流	步骤固定、可验收	灵活性差
多 Agent	需要分工、并行、专业工具	调试与成本高

先用工作流把流程跑通，再决定要不要“智能”。多数业务不需要多 Agent。

四、边界写在哪

- 系统提示：能做什么、不能做什么、缺信息怎么办；
- 工具层：只暴露必要接口，危险动作单独授权；
- 编排层：最大步数、超时、成本上限；
- 审核层：哪些动作必须人工点确认；
- 日志层：每次调用可回溯。

五处都写，才算真的“有边界”。

五、30 分钟练习

1. 选一个你熟悉的流程（5 分钟）；
2. 画出五层架构（10 分钟）；
3. 标出四个“不该自动”的动作（5 分钟）；
4. 写三条系统提示里的禁止项（5 分钟）；
5. 选一种架构并说明理由（5 分钟）。

六、验收清单

- [ ] 五层架构齐全，审核层存在；
- [ ] 不可逆动作全部有人工确认；
- [ ] 系统提示写清“缺信息怎么办”；
- [ ] 有最大步数与成本上限；
- [ ] 架构选择有理由，不是跟风。

七、延伸阅读

- 通识版第 7 讲《认识智能体》
- 知识库 `1.3 AI Agents（智能体）`

> 下一讲：工具调用与记忆。架构定了，接下来让 Agent 真正“能动手、记得住”。