M3 · 职业版 Agent · 第 4 讲

第 4 讲：常驻 Agent 与成本

事件驱动，不是一直跑

阅读约 25 分钟

练习约 30 分钟

产出：成本表 + 预算上限

本节目录

一个真实故障

本讲目标

一、五个部件

二、去重与幂等

三、成本怎么算

四、预算上限

五、该看的指标

六、30 分钟练习

七、验收清单

八、延伸阅读

一个真实故障

团队把 Agent 设成每分钟检查一次邮箱。周末没人看，它反复处理同一封邮件，两天烧掉一个月的预算。

常驻不等于“一直跑”。真正省钱的常驻，是**事件驱动 + 去重 + 预算上限**。

本讲目标

- 用五个部件搭一个常驻循环；
- 用队列去重与幂等避免重复劳动；
- 算出单任务成本并设上限；
- 知道该看哪几个指标。

一、五个部件

部件	作用
收件箱	接收任务，排队而不是直接执行
计时器	定时触发、延迟重试
记忆	记住进度，避免重复
权限	能碰哪些系统
发件箱	产出通知、写回业务系统

这五个部件来自 Anthropic 的在线 Agent 架构实践：把“持续运行”拆成可观察的队列与状态。

二、去重与幂等

- 每个任务带唯一 id（邮件 id、工单号）；
- 处理前先查“是否已处理”；
- 同一 id 的结果写回同一记录；

- 失败重试不超过 2 次，然后进“待人工”队列。

没有唯一 id 的常驻任务，等于定时制造重复数据。

三、成本怎么算

项	计算
输入 token	上下文 + 材料
输出 token	回答 + 工具参数
工具调用	按次计费或按 API 价格
存储	日志与附件

单任务成本 = 一次运行的全部消耗。先测 20 次取平均，再乘每日任务量。

四、预算上限

- 单任务上限：超过即中止并转人工；
- 每日上限：达到后只处理高优先级；
- 每月上限：按周监控，80% 时预警；
- 异常检测：单任务成本超过均值 3 倍，自动暂停。

把上限写进编排层，不要只写在文档里。

五、该看的指标

指标	说明
队列长度	积压说明处理不过来
重复率	同一 id 被处理多次
失败率	按工具与错误类型拆分
单任务成本	按任务类型统计
人工介入率	太高说明能力或规则不足

六、30 分钟练习

1. 选一个每日任务（5 分钟）；
2. 画出五部件与数据流（10 分钟）；
3. 设计唯一 id 与去重规则（5 分钟）；
4. 估算单任务成本与日预算（5 分钟）；
5. 写三条告警阈值（5 分钟）。

七、验收清单

- [ ] 任务走队列，不是定时直接执行；
- [ ] 每个任务有唯一 id 与去重检查；
- [ ] 重试有上限，失败进人工队列；
- [ ] 单任务/每日/每月都有预算上限；
- [ ] 五个指标有看板或日志。

八、延伸阅读

- 参考文章：Anthropic 7×24 在线 Agent 架构指南
- 知识库 `1.3 AI Agents（智能体）`

> 下一讲：企业级 Agent OS。单个 Agent 跑通后，才谈多团队共享。