M3 · 职业版 Agent · 第 3 讲

第 3 讲：评测与验证

把“做了没验”变成硬步骤

阅读约 25 分钟

练习约 30 分钟

产出：20 条评测集

本节目录

一个真实故障

本讲目标

一、三层验证

二、把标准写成断言

三、20 条评测集怎么配

四、验证循环

五、停止条件

六、指标怎么看

七、30 分钟练习

八、验收清单

九、延伸阅读

一个真实故障

Agent 改完代码，回复“已完成，测试通过”。你打开一看：测试根本没跑，它只是“觉得”没问题。第二天上线，接口 500。

Agent 最常见的失败不是不会做，而是**做了没验**。这一讲把“验证”变成流程里的硬步骤。

本讲目标

- 把验收标准写成可执行检查；
- 建一套 20 条的评测集；
- 设计“检查—诊断—重试—升级”的循环；
- 定停止条件，避免无限重试与成本失控。

一、三层验证

层	做什么	谁来做
自动检查	跑测试、校验格式、比对字段	程序
评测集	20 条真实任务，统计通过率	程序 + 人
人工抽检	看语气、判断、边界	人

只有第一层，容易“跑通但不对”；只有第三层，规模上不去。

二、把标准写成断言

“回答要准确”不可执行。改成：

- 输出必须是合法 JSON；
- 字段必须包含 id、category、reason；
- category 只能是四个枚举值；
- 引用必须能在原文中找到；
- 回答长度不超过 200 字。

每条都能写成一行检查。写不出检查的标准，先不算标准。

三、20 条评测集怎么配

类型	数量	例子
正常	8	标准输入、常见任务
边界	6	空值、超长、多语言
失败	4	缺字段、权限不足、超时
对抗	2	诱导越权、提示注入

每条记录：输入、期望、判定方式、实际结果。评测集要版本化，改一次记一次。

四、验证循环

1. **执行**：Agent 产出结果；
2. **检查**：跑断言与测试；
3. **诊断**：失败归因（输入错/工具错/逻辑错）；
4. **重试**：只重试可恢复错误，最多 2 次；
5. **升级**：仍失败则转人工，附日志。

重试不是“再跑一遍”，而是带着诊断信息修一处再跑。

五、停止条件

必须在系统提示或编排层写死：

- 最大步数（如 15 步）；
- 最大时长（如 5 分钟）；
- 单任务成本上限；
- 连续失败 2 次即停；
- 涉及付款、删除、外发，必须人工确认。

没有停止条件的 Agent，迟早会烧掉一笔预算。

六、指标怎么看

指标	健康值	说明
通过率	⩾ 80%	低于就先别上线
一次成功率	⩾ 60%	反映稳定性
平均步数	稳定	突然上升说明在瞎试
人工介入率	可解释	太高说明能力不足
单任务成本	有上限	按周统计

七、30 分钟练习

1. 为一个任务写 5 条可执行断言（10 分钟）；
2. 配 10 条评测（正常 4、边界 3、失败 2、对抗 1）（10 分钟）；
3. 画出验证循环与停止条件（5 分钟）；
4. 跑一遍，记录通过率与失败归因（5 分钟）。

八、验收清单

- [ ] 每条标准都能被程序检查；
- [ ] 评测集覆盖四类输入；
- [ ] 重试有上限，且带诊断；
- [ ] 有转人工路径；
- [ ] 成本与步数有硬上限；
- [ ] 评测结果有版本记录。

九、延伸阅读

- 知识库 `1.3 AI Agents（智能体）`
- 知识库 `4.3 团队搭建：AI 招聘/组队`
- 参考文章：Claude Code 团队如何思考验证；Anthropic 7×24 在线 Agent 架构指南

> 下一讲：常驻 Agent 与成本控制。验证过关后，才谈 7×24 运行。
