A FIELD NOTE ON AI DATA & LEARNING SYSTEMS

当模型不再缺答案，什么数据还值钱？

从“内容库存”到“可学习的工作过程”
企业 AI 的下一道护城河，是把每一次运行变成下一次进步。

本文提出一套独立框架：轨迹 TRACE · 评测 EVAL · 治理 GOVERNANCE。
它们共同决定企业能否把 Agent 从一次演示，变成持续工作的系统。

九天

2026.08 · 基于公开资料与业务观察

00 EXECUTIVE SUMMARY

真正稀缺的，不是更多内容

数据的价值，正在从“你有什么”，
转向“你的组织怎样把事情做成”。

1 模型不缺常识，缺业务边界

公开知识越来越容易获得，真正难的是在具体权限、规则和目标下做对判断。

2 结果不够，过程才解释原因

最终答案只能告诉我们成败；意图、动作、异常与修正，才告诉系统下次怎样做得更好。

3 真实轨迹不是自动奖励

销量、项目成功或客户采用都受多重因素影响，还需要黄金任务、评分标准与人工复核。

4 稳定预算通常发生在上线以后

评测维护、失败分析、权限治理、回归测试和人审运营，都会随着 Agent 长期运行而持续发生。

5 护城河是学习速度，而不是数据堆积

谁能更快发现失败、解释失败、修正系统并安全发布，谁就拥有复利。

63%

Gartner 对 1,203 名数据管理负责人调查：63% 的组织没有，或不确定自己是否具备适用于 AI 的数据管理实践。

01 FROM DOCUMENTS TO DECISIONS

为什么“多给它一些资料”很快失效

传统知识库解决的是“去哪里找到答案”。当 Agent 开始替人行动，问题变成“此刻应该做什么、做到哪一步、什么时候停下来交给人”。后者需要的不是更多段落，而是组织真实做事时留下的判断证据。

传统数据项目

收集文档 → 清洗 → 标注 → 交付

它最擅长：

- **提供背景知识**
- **统一术语与口径**
- **支持检索与问答**

盲点：文档通常记录“组织说过什么”，很少记录“组织为什么这样决定”。

Agent 运行项目

接到任务 → 使用工具 → 遇到异常 → 修正 → 交付

它必须理解：

- **目标与优先级**
- **权限与风险边界**
- **例外、升级和停手机制**

关键：让系统学习“在约束下完成任务”，而不只是生成看起来合理的文字。

一个实用判断：
如果数据只能回答“正确答案是什么”，它更像知识；如果还能回答“为什么这样做、哪里出错、谁修正、结果怎样”，它才开始成为可学习的业务过程。

02 THE TRACE MODEL

一条真正有用的工作轨迹，至少有六层

不是把屏幕录像或聊天记录一股脑存下来，而是把一次任务拆成可解释、可授权、可评测的结构。

01

任务意图

为什么做、要解决什么、成功长什么样。

02

上下文与约束

规则、权限、时限、预算、禁区与依赖。

03

动作与工具

查了什么、调用了谁、采取了哪些步骤。

04

异常与歧义

信息缺失、冲突、失败、意外和不确定性。

05

人类修正

哪里被改、为什么改、谁有最终决定权。

06

结果与证据

交付是否被采用、质量如何、证据在哪里。

例子：供应商筛选

低价值记录只有“已完成筛选”；高价值轨迹还包括候选来源、淘汰理由、风险项、缺失证据、谁做了升级判断，以及最终决策是否经得住复核。

03 REALITY NEEDS EVALUATION

最容易误判：有真实结果，就有可靠答案

真实业务数据解决的是分布问题：它让我们看到事情在现实中怎样发生。但它不会自动解决因果归因，也不会自动告诉模型哪一步值得学习。

外部变量

销量受价格、季节、渠道和竞争影响，不能都归因给 Agent。

幸存者偏差

只保存成功案例，会把运气和不规范操作一起学进去。

延迟反馈

关键动作今天发生，真实后果可能数周后才出现。

隐性知识

专家凭经验停手或升级，但日志里可能只留下一个点击。

把真实轨迹变成可学习信号，需要一层“评测翻译”

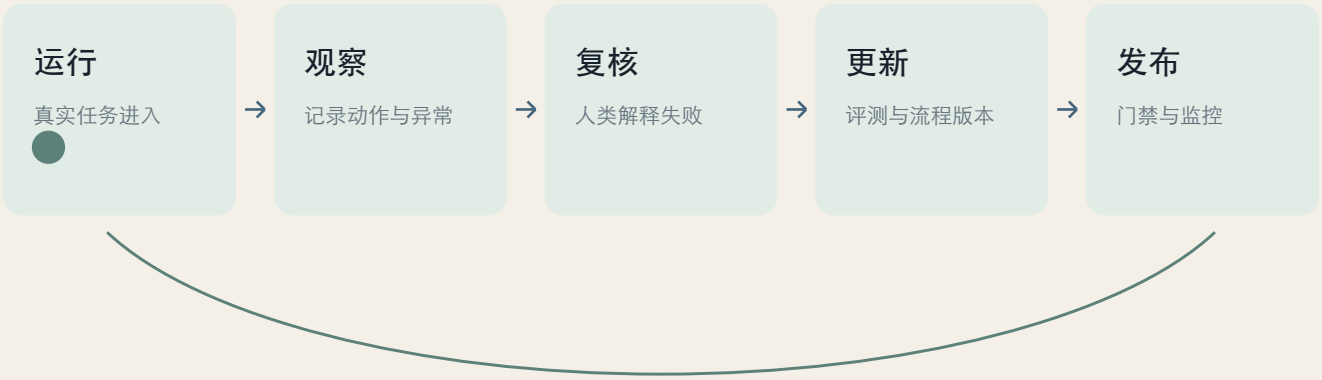
真实任务（保留上下文） → 黄金样本（覆盖关键边界） → 评分标准（解释好坏） → 人工复核（处理灰区） → 业务指标（验证价值）

轨迹负责把现实带进系统；评测负责决定系统应该从现实中学什么。

04 THE PRODUCTION LEARNING LOOP

企业真正会长期付费的，是上线后的改进

一次数据迁移、一次 Prompt 调优、一次演示环境，都可能结束。只要 Agent 真的进入生产，新的失败、规则和权限就会持续出现，改进工作也就不会停止。



容易一次性结束的预算

- **数据迁移与初始清洗**
- **首版知识库和 Prompt**
- **演示用任务环境**
- **一次性的模型微调**

交付物做完，预算就可能结束。

会随运行持续发生的预算

- **评测集与规则维护**
- **失败分析和人工复核**
- **权限、审计与合规门禁**
- **换模型、换工具后的回归测试**

价值不在“再做一批数据”，而在“让系统持续可靠”。

商业含义：比起按条数、人头或标注小时收费，更值得设计的是“首期诊断与实施 + 持续评测运营 + 可选结果变量费用”。

05 POLICY IS NOT GOVERNANCE

别把模型厂商想成会自动吸走一切的黑箱

“企业越用模型，经验就越自动变成厂商训练数据”这个说法并不准确。至少在当前公开政策下，主流商业产品对客户数据的默认使用有明确边界。

OpenAI 商业产品与 API

官方说明：默认不使用组织在 ChatGPT Business、Enterprise、Edu 及 API 等产品中的输入或输出训练模型；明确选择加入时除外。

Anthropic 商业产品

官方说明：不会使用商业客户的聊天或编码会话训练模型，除非客户加入相关计划、明确反馈或主动选择。

真正需要企业自己守住的，是下面五件事

- **权限：谁能看、谁能调用、谁能批准高风险动作。**
- **保留：数据保留多久，日志能否删除，是否需要本地化。**
- **连接器：第三方工具、插件和外部系统会把什么数据带出去。**
- **资产化：Prompt、Skill、流程和评测是否沉淀在企业可控的系统里。**
- **退出：换模型或换供应商时，能力能否迁移，历史记录能否带走。**

注：公开产品政策不替代合同审查、权限设计和具体法域的合规判断。

06 FIVE COMPOUNDING ASSETS

真正能积累的，不是原始数据堆

可持续 AI 能力 = 业务语义 × 可验证轨迹 × 迭代速度 × 权利边界

01	业务对象模型	客户、合同、任务、风险、审批和结果之间的关系。
02	黄金任务与评测集	真正代表业务边界、可以反复回归测试的样本。
03	错误分类与人工修正	失败为什么发生，专家怎样纠正，下一版如何避免。
04	运行版本账本	模型、Prompt、Skill、Workflow、工具和评测的版本关系。
05	权利矩阵与审计记录	什么可训练、可评测、可保存、可跨域使用，谁批准。

跨客户真正适合复用的，是对象模型、评测方法、错误分类、流程模板和安全门禁；客户原始内容默认留在客户域内。

07 A 90-DAY FIELD PLAN

先跑通一个闭环，再谈平台化

选择一个高频、可量化、可人工复核的窄流程。目标不是做出“万能 Agent”，而是证明失败可以被发现、解释、修正和复测。

1

第 1-2 周 · 选流程

明确负责人、成功标准、禁区与当前基线。

2

第 3-4 周 · 建结构

拆任务对象、权限、数据来源和升级路径。

3

第 5-8 周 · 采轨迹

记录动作与异常，建立首批黄金任务和人审标准。

4

第 9-12 周 · 影子运行

回归测试、发布门禁、小流量试点和业务复盘。

90 天验收：系统能解释主要失败类型；每次改动都有版本；发布前能回归测试；高风险动作有人工批准和回滚。

08 MEASURE THE SYSTEM, NOT THE DEMO

别只看“准确率”，要看四组指标

Agent 的表现是模型、工具、权限、流程和人的共同结果。只有同时看质量、效率、业务和控制，才知道系统是否真正进步。

质量

任务成功率
关键事实错误率
严重风险漏检率

效率

人审分钟数
单任务综合成本
平均处理时长

业务

周期缩短
返工减少
采用率与完成率

控制

越权次数
审计完整性
回滚与恢复时间

每次升级都要回答三个问题：**改了什么？为什么更好？出了问题怎样退回去？**

09 THE INVESTMENT & PRODUCT CHECKLIST

判断一个 AI 数据机会，先问这八个问题

1. 谁拥有数据，谁明确授权了训练、评测、保存和复用？
2. 数据会持续产生，还是交付一次就开始过期？
3. 结果能否被独立验证，还是只能靠主观感觉？
4. 失败能否归因到具体步骤、规则、工具或权限？
5. 人工修正能否结构化，进入下一轮评测？
6. 客户上线以后，是否仍需要监控、复测与治理？
7. 可复用的是方法和结构，还是依赖复制客户原始内容？
8. 换模型、换供应商、换工具后，能力还能否迁移？

如果前六个问题答不清，先别急着谈规模。
真正可持续的产品，不是把数据卖出去，而是让客户在每一次真实运行中保留判断、发现失败，并更快变好。

SOURCES & NOTES

资料、口径与原创说明

本文为九天基于公开资料与企业 AI 项目观察形成的独立分析框架。为避免用未经证实的市场传闻制造确定性，正文没有采用私营公司未审计的预算、市场份额和估值倍数。

Mercor	2026-05-08 官方披露：年初超过 10 亿美元 annualized revenue run rate，并向 3 万多名活跃承包者每天支付超过 200 万美元。run rate 不等于 ARR、净收入或未来合同。 https://www.mercor.com/blog/when-you-go-from-2-million-a-month-to-2-million-a-day/
Gartner	2025-02-26：对 1,203 名数据管理负责人的调查显示，63% 的组织没有或不确定自己是否具备适用于 AI 的数据管理实践；AI-ready data 不是一次性项目。 https://www.gartner.com/en/newsroom/press-releases/2025-02-26-lack-of-ai-ready-data-puts-ai-projects-at-risk
NIST AI RMF	生成式 AI 风险管理框架强调在 AI 生命周期中持续治理、映射、测量和管理风险，并将测试、评估、验证与确认作为重要能力。 https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence
OpenAI	商业数据政策：组织在 ChatGPT Business、Enterprise、Edu 及 API 等产品中的输入和输出默认不用于训练或改进模型，明确选择加入时除外。 https://openai.com/business-data/
Anthropic	商业产品政策：不会使用商业客户的聊天或编码会话训练模型，除非客户明确选择相关计划或提供反馈。 https://privacy.claude.com/en/articles/7996885-how-do-you-use-personal-data-in-model-training
国家网信办	数据跨境和个人信息处理需结合数据类型、主体角色、数量阈值及具体业务场景判断。本文不构成法律意见。 https://www.cac.gov.cn/2026-07/24/c_1786638883119336.htm

© 2026 九天。可私下转发；公开转载或商用请注明作者。